TWO ASPECTS OF COMPUTERIZED PATIENT HISTORY THAT IMPROVE BEDSIDE PULMONARY DIAGNOSIS

by

Tracy Rich

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Medical Informatics

The Univeristy of Utah

March 1989

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a thesis submitted by

Tracy Rich

This thesis has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

Chair: Peter J.

Pryor

Homer R. Warner

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the thesis of _____ in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; (3) the final manuscript is satisfactory to the Supervisory Committee and is ready for submission to the Graduate School.

_____ _____

Date

Approved for the Major Department

_____

Homer R. Warner

Chair/Dean

Approved for the Graduate Council

_____

B. Gale Dick

Dean of The Graduate School

ABSTRACT

This thesis project involves two aspects of research in the area of diagnostic decision making of cardio-pulmonary diseases. This study is concerned with a set of logic frames that have been written to detect a group of 41 pulmonary and cardiology diseases. When the diagnostic frames were first developed, there were 58 frames for 29 diseases. There were two frames for each disease. The first dealt with patient history, and the second dealt with physical exam, laboratory, and radiology findings. The a posteriori probabilities generated by the history frames are used as the entrant a priori probabilities of the corresponding clinical data frames. The history frames have evolved to include more than the original 29 disease states, whereas the clinical data frames have continued to include only the original 29 diseases.

The first part of the study deals with the problem of gathering a computerized patient history that could then be available for use by the pulmonary diagnostic history frames. Two different computerized data gathering techniques, that captured a patient history directly from the patient using a terminal mounted near the bedside, were compared. One of the collection modes is a hierarchical questionnaire that consists of a conditional branching set of fixed frames (GQAP). The other history gathering program utilizes DDA, a decision-driven data acquisition system. The goal of this project is to determine if there is a measurable difference between the two data collection modes in 1) time to complete, 2) number of questions asked, or 3) accuracy of decision made from the entered data.

Physical exam data entered by the physician into a computer terminal were the next type of data that was scheduled to be collected at the patient's bedside. These data would then be available for utilization by the clinical data frames that diagnose cardio-pulmonary diseases. Before the collection of physical exams began, the Bayesian statistics used in the clinical data frames for chest X-ray findings needed to be revised. It was felt that revising the statistics for the radiology finding, using a data base of actual patients who had entered LDS Hospital, would create a better set of statistics than the statistics that were originally estimated by the medical experts. The second part of this study involves the revision of

chest X-ray findings in the pulmonary clinical data frames, and the evaluation of the effect the revised statistics had upon the accuracy of the diagnostic system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The focus of this research project is twofold, the first being the comparison of two methods used for collection of computerized patient histories and their ability to generate an accurate differential diagnostic list of diseases. The second focus involved using statistics derived from a clinical data base to revise the diagnostic logic of the Bayesian system used by the HELP system for generating a differential diagnostic list of pulmonary diseases based upon the physical exam findings.

Research Goals for Comparison of History Collection Methods

Since a comprehensive and accurate history is essential to good medical practice through its role of understanding and management of patient care (1, 2), two history collection modes were compared to determine if one would be better (faster, more accurate, or enabled more accurate decisions to be made based on the entered data) to use as a tool for gathering a patient history. A hierarchical branching questionnaire was compared to a decision-driven data acquisition system, both collecting histories from patient with pulmonary diseases. The goal of this part of the research project was to determine if there is a difference between the two collection modes based on 1) time to complete, 2) number of questions asked, or 3) accuracy of decisions made from the entered data. This was accomplished by presenting a selected population of pulmonary patients with a computer terminal that collected their histories using either the hierarchical branching questionnaire or the decision-driven data acquisition system.

Advantages and Benefits of Computerized Patient History Collection

A computerized method for the purpose of collecting patient histories has many advantages over the traditional physician administered interview. The manual methods can omit a large amount of relevant information, consume time that can be used by the physician for patient counseling and produce a history that is often illegible and badly structured. Computerized medical histories present a uniform and consistent line of

questioning, save physician time and produce a legibly retrievable record of patient responses (1, 2, 3, 4, 5).

The computerized patient history was developed with the concept that it could be a very useful tool for physicians. A study preformed by Teach et al. (6) indicated that physicians are accepting of applications that enhance their patient management capabilities, but tend to oppose applications in which they perceive an infringement on their management roles. While automated history-taking systems have not yet received widespread acceptance, it is hoped that with exposure and the assurance that they will provide a invaluable tool to the physician, they will become an important diagnostic aide.

Computers can usefully supplement the decision making process through their ability to refer to a massive database of information and reiterate faithfully every item stored in it. This should be viewed as a great help to the physician based on the distinct possibility that one physician cannot know everything or have complete and unbiased experience of everything in his specialty (7).

Since there are limits to human capabilities as an information processor, the occurrence of random errors in their activities is assured. Therefore many medical errors are probably due to the physicians' intrinsic limits rather than to remedial flaws in their fund of knowledge. These errors have been shown to be reduced by the interaction of physicians with computer generated suggestions about the management of simple clinical events (8, 9). Not only has the computer-aided diagnosis been shown to be a direct adjunct to the clinicians, they may also benefit in the short-term from the constant feedback received from the disciplines and constraints involved in communicating with the computer (10).

Review of Computer Assisted History Collection

Following is a brief review of computerized systems that have been developed in the past to aide in the collection of patient medical history.

One of the first computer-based medical-history systems in which the computer collects information directly from patients was developed by Slack and his co-workers at the University of Wisconsin in the mid 1960s (4). The questionnaire they developed was branching in nature, where the choice of questions presented and the order of their presentation are a function of the patient's responses. A "Yes" response to a general question is followed by a series of specific qualifying question. A "No" response to a general question results in the skipping of unnecessary qualifying questions and the presentation of another general question. The questionnaire dealt with gathering symptoms of allergy from a patient sitting in front of a CRT and responding to the questions presented

on the screen by pressing numbered keys on the keypad. The questionnaire contained 500 questions, with 29 the minimum asked and 320 the maximum. Once the history was concluded, the patient's responses were printed on paper for immediate clinical use.

The Mayo Clinic developed a computer-input system that employs video screen and light-pen to allow the physician to enter and retrieve patient historical and physical-examination data (11). In operation this video-matrix system initially displays a video screen containing general categories of information. As a selection is made, succeeding displays first automatically narrow the scope and then sequence the selections from general to specific until the desired entry data are available. Use of this system resulted in clinical records that were carefully formatted, complete and legible, and easily retrievable either on the video screen or as printed records.

A computer program that uses unstructured problem-solving technics was developed by Pauker et al. (12) to take the history of present illness of patients with edema. This program tries to develop a sufficient understanding of the patient's complaints to form a reasonable basis on which to evaluate the clinical problem and to lay to ground-work for subsequent management decisions. It accomplishes this by using processes of both information-gathering and diagnosis. The program alternates between asking questions to gain new information and integrating this new information into the developing picture of the patient. This program uses "frames" (information about a disease) in its attempt to place the patient within a disease classification. These frames are related to specific diseases and to various clinical and physiologic states associated with these diseases. The frames contain descriptions of typical findings, numerical factors to be used in scoring, and links to other frames. The program cycles through characterizing findings, seeking advice on how to proceed, generating hypotheses, testing hypotheses and selecting questions until all active hypotheses are explored.

Lilford and associates designed a patient-interactive microcomputer system to obtain histories from patients attending an infertility and gynecologic endocrinology clinic (1, 13). The patients enter their answers directly into the system through the use of a keyboard containing "yes', "no", and "don't know" buttons and the numbers 1 to 5. The history follows a highly branching pattern, whereby positive responses lead to more detailed questions on that topic. The program contains 342 questions with an average of 88 questions asked during the course of an average history collection session (13). The system also provides a neatly formatted summary of the history to be printed out for the use of physicians. Their study showed that the computer system can generate considerably more information than a conventional antenatal history thereby providing a much more complete history.

Research Goals for Revision of Diagnostic Logic

The HELP hospital information system, in order to provide on-line, real-time, automated data interpretation and alerting capabilities, has a medical decision module built into the system. This subsystem is driven by medical knowledge that is stored as a set of knowledge-description frames. This system has been used to produce a list of differential diagnoses early in the medical examination and treatment process. The approach is to generate a one to five member differential diagnostic list using only the patient data available within the first day of hospitalization, which can be charted for physician review. Data from the patient history, the admitting laboratory studies, and the initial chest X-ray are included in the decision frames. This system uses a Bayesian model of diagnosis.

In order to proceed with the next phase of this project, the entry of physical exam data into the bedside computer system by the physician, it was felt that the statistics in the original pulmonary physical exam frames needed to be updated. When these diagnostic frames were originally designed a group of medical experts estimated a priori probabilities for each disease and a sensitivity and specificity for each manifestation within the disease frame. These statistics used by the diagnostic logic were revised through the use of a clinical data base. The revised statistics were evaluated to determine if an improvement over the originally estimated statistics had been achieved. The evaluation of the two sets of statistics (estimated and revised) includes a comparison of sensitivities and specificities within each group, and a comparison of the accuracy of the differential diagnostic lists produced using the using the original frames and those produced using the revised frames.

Review of Medical Decision-Making Systems

Expert systems, computer programs capable of performing a task that normally requires the knowledge of an expert in that field, have been developed over the past few decades in the area of medical decision making. These computer-assisted diagnostic systems have been developed in order to assist (not replace) the physicians in making a diagnosis. The rationale for using computers is that it is impossible for one doctor to know everything or even to have complete and unbiased experience of everything in their specialty (7). If dependable models for disease diagnosis could be constructed, they could would have a vast impact on the costs and availability of health care. Diagnostic clinics, located in areas where health care availability is limited, manned chiefly by paramedical personnel using automated diagnostic models, could process thousands of patients per year at a low per patient cost. The diagnostic models could also be used to calculate the relative probabilities of the presence of various diseases in cases where the relationship between symptoms and diseases is so complex that no clear-cut diagnosis can be made (14).

There are two main approaches to computer assisted medical decision making. The first is a mathematical process in which calculated probabilities are used to describe a condition. The second consists of a system of rules known as the knowledge base, which forms the basis on which the computer can give an analysis (7). Croft (14) concluded in his research that the various mathematical-diagnostic models that have been tested in the past differ very little from one another in terms of their theoretical underpinnings and the diagnostic results they produce. He felt that automated diagnosis was possible but that future developments should be more concerned with resolving 1) lack of standard medical definitions, 2) lack of large, reliable medical data bases and 3) lack of acceptance of computerized diagnostic aids by medical professionals rather than with the construction of new mathematical models.

The following is a brief review of some of the medical decision making systems that have been developed over the past few decades.

Gorry and Barnett in the late 1960s used Bayes rule to develop a model of sequential diagnosis (15). They believed that diagnosis consists of two major functions, inference and test selection. Most of the programs developed before this time dealt with diagnosis through inference only. They wanted to take diagnosis one step further by developing a program that was able to also determine an appropriate sequence of diagnostic test to perform on the patient. The program operates in an interactive mode. The basic mode of inference employed by the program is probabilistic, and a major portion of the information requirement consists of probabilities. The program employes a Bayesian analysis of attributes as its central inference mechanism with the current view held by the program being a conditional probability distribution for the various diseases.

The program basically operates in the following manner. The physician defines a problem for the program by indicating some set of initial attributes, from which the program obtains a current distribution. At this point the program either selects a test to be performed or it makes a final diagnosis and ceases testing. The program employees a decision-tree analysis in the test selection function. Each test is weighed according to cost, expected outcome, and relative importance in avoiding a misdiagnosis. The test selection function will continue to select tests to be performed as long as at least one test is expected to reduce the decision loss by an amount exceeding the cost of the test. Only when no test is expected to reduce the decision loss sufficiently to justify its expense does the function make a terminal decision.

MYCIN, developed at Stanford by Shortliffe and associates, is an interactive computer program that uses the clinical decision criteria of experts to advise physicians regarding the selection of appropriate antimicrobial therapy for hospital patients with

bacterial infections (16). MYCIN gives advice in this area by means of three subprograms: 1) A consultation system that uses information provides by the physician, together with its own knowledge base, to choose an appropriate drug or combination of drugs; 2) An explanation system that understands simple English questions and answers them in order to justify its decisions or instruct the user; and 3) A rule acquisition system that acquires decisions criteria during interactions with an expert and codes them for use during future consultation sessions.

All knowledge used by MYCIN during a consultation session is contained in therapeutic decision rules. Each rule consists of a set of preconditions (PREMISE), which if true, permits a conclusion to be made or an action to be taken, according to the ACTION part of the rule. A goal-oriented control structure allows MYCIN to select appropriate rules and ignore those that are not applicable to the current patient. As soon as a condition within the PREMISE of a rule is found to be false, that rule is rejected. If a condition is found to be missing, FINDOUT either derives the necessary information from other rules or asks the user for the data.

HEME (17) is a computer program that was conceived and developed to provide physicians with diagnosis-oriented analysis for hematologic diseases. Its uses include suggesting diagnoses, reminding doctors of available testing procedures, and being able to check physicians' thinking at each stage of the diagnostic process. HEME is a Bayesian program that is different from other Bayesian programs by the fact that each disease is analyzed individually in order to determine the probability that the patients have the disease versus the probability that they do not. This allows for the possibility that a patient could have more than one disease. HEME also contains a feature that automatically improves its diagnostic accuracy by updating the probabilities used in Bayes' theorem as clinical data accumulates in the data base.

SPHINX (18) is a system that deals with the relatively unbound problem of diagnosis in internal medicine. It is based on methods of inference and pattern matching and on various heuristic features. The system receives information in two ways: data provided by the patients about their complaints; and data requested by the system itself. The decision modules contain two components: a set of rules, each representing a part of clinical medical knowledge; and a control structure that activates and selects the rules. This is a pattern-directed inference system. It is a data-driven program that is able to react to each change in its environment.

The patterns are organized as tree structures with true or false rules attached to each node that include complementary signs and data that must or can by required for the accuracy of a diagnosis. The control is explicitly carried out by means of several heuristic

features that take into account costs of examinations and successful branching rate. The processing of SPHINX is composed of three steps: 1) Establish the patient's context from the signs that led to the consultation; 2) Determine the set of syndromes of which the signs presented by the patient could be manifestations and determine the range of the various diagnostic orientations; 3) Require complementary investigations where they are necessary to identify the suggested diseases.

INTERNIST-I is an experimental computer program developed by Miller and associates, which is capable of making multiple and complex diagnoses in internal medicine (19). Using a patient's initial history, physical exam findings, and laboratory results it aids the physician with making multiple and complex diagnoses. The system derives its diagnostic capabilities from its extensive knowledge base and from heuristic computer programs that can construct and resolve differential diagnoses. INTERNIST-I is an example of applied symbolic reasoning that represents an attempt to model the behavior of physicians.

The basic structure of the knowledge base is the individual disease profile. These list findings (historical items, symptoms, physical signs, and laboratory abnormalities) that can occur in patients with each illness. The individual diseases are part of a disease hierarchy that is organized from the general to the specific. The knowledge bases also detail relations among diagnoses and among manifestations. The behavior of INTERNIST-I results primarily from the application of two heuristic principles: the formation of problem areas through a partitioning algorithm and the conclusion of diagnoses within problem areas, using strategies such as diagnosis by exclusion.

INTERNIST-I uses the following steps during a diagnostic consultation. Initial findings of the patient are entered by the user. The program creates its complete differential diagnosis from the inverted disease profiles. A disease hypothesis is created for each item on the manifestation's differential diagnosis list. Each hypothesis on the master list of diagnoses is given a score. The master list of all hypotheses is then sorted by descending score. Scores are determined by the presence or absence of a manifestation required in the hypothesis. A problem area containing the topmost disease hypothesis is created. If there is more than one diagnosis in the problem area, additional questioning takes place to resolve the diagnostic decision. Disease hypotheses are scored again after each group of questions are answered. The program continues to cycle (questioning, scoring hypotheses, and selection problem areas) until there is only one disease in the problem area. This disease is considered the patient's diagnosis. If a single diagnosis cannot be resolved, the program presents a list of all hypotheses in the top problem area, in order of descending score.

This research project employed the HELP hospital information system, developed by Warner et al., which uses a Bayesian method of classification (20). This system is explained in more detail the Methods section.

METHODS

This thesis covers two aspects involved in the acquisition and manipulation of medical data for the purpose of diagnostic decision making. The first concerns the collection of a computerized patient history; the other is the revision of diagnostic logic. This chapter includes a description of how two modes of patient history collection were implemented and compared, and a description of how the statistics for radiology findings that are used in pulmonary disease logic sectors were revised through the use of a current patient data base.

Methods for Comparison of Two Modes of Data Collection Used to Acquire Patient Histories

Gathering qualitative data, such as a patient history, from human sources remains a major problem in medical information systems. In this study we compare two different computerized data gathering techniques that capture patient history directly from the patient, using a terminal mounted near the bedside. The goal of this study is to determine if there is a measurable difference in 1) time to complete, 2) number of questions asked, or 3) accuracy of decisions made from the entered data, using a conditional branching set of fixed frames (GQAP) as opposed to a decision-driven data acquisition (DDA) system (21).

This section contains a description of the implementation of the patient history collection modes within in the context of the HELP system using subsystem tools (PTXT, QSTN, and HCOM), a description of each of the patient history collection programs (GQAP, a branching questionnaire entry system, Decision-driven data acquisition (DDA)), a description of the method used to collect the history from the designated patient population used in this study, and a description of the methods used for comparison of the two different modes of patient history collection.

Use of the HELP System

HCOM

The Health Evaluation Through Logical Processing (HELP) system is a comprehensive computer system used for acquiring medical data and implementing medical decision logic. It is not only a traditional Hospital Information System (HIS), but is also capable of using knowledge based decision making applications for clinical care. The goal of this system has been to improve the delivery of medical care by increasing the availability of medical data to health care professionals (21).

The most outstanding aspect, and one of particular interest to this thesis, of the HELP system is the ability to allow construction of modular sectors that may be used in the medical decision-making process. The HELP programs used for developing this medical logic were designed to be "user-friendly," allowing experts in the subject matter to design sectors for specific decisions without formal training in programming. This means that each decision module should be readily understood by physicians and other members of the medical community who are not trained in computer science. Each HELP sector is user defined and can be as simple or complex as needed to solve the medical problem at hand. Sector logic can consist of straightforward deterministic "if . . . , then . . . " rules or probabilistic applications based upon the Bayes formula (22). Sectors access raw data stored on the patient file, and use that data to create and store back out on the patient's file, the medical decision for which that particular sector was written.

This study was concerned with a set of sectors that were written to detect a group of 41 pulmonary and cardiology diseases (23) (Table 1). After the patient history was collected, this pulmonary disease knowledge base was used to create a differential diagnostic list of pulmonary diseases for a specific patient.

Following is a description of the main areas of a HELP sector. Then to illustrate how these sectors are written and how they work, an example of a sector used to predict the presence of bacterial pneumonia will be explained.

Each sector has the following components: 1) a text string or title, 2) a list of destinations to which the text is to be sent, 3) a list of data items from the HELP dictionary that is to be used in making the decision, and 4) the logical expressions that define the decision criteria (24). The data items in a sector are expressed with text that must be explicitly defined in the PTXT file, the dictionary of the HELP system. The HELP sector text, or message itself, becomes part of the PTXT file.

HELP sectors are organized into a list of items that are labeled alphabetically. Three basic item types are used in building sector logic. An item may specify a search to be performed on the patient's data file (search item), an arithmetic or logical relation among

Table 1. List of 41 diseases used in this research

Acute Bacterial Bronchitis	Legionnaires Disease
Acute Pericarditis	Lung Abscess
Angina Pectoris	Metastatic Neoplasm
Asbestosis	Mycoplasma Pneumonia
Aspiration Pneumonia	Myocardial Infarct
Asthma	Non-Hodgkin's Lymphoma
Bacterial Pneumonia	No Pulmonary Disease
Bronchiectasis	Pneumococcal Pneumonia
Chronic Bronchitis	Pulmonary Coccidiomycosis
Chronic Pulmonary Histoplasmosis	Pulmonary Edema/CHF
Coal Worker's Pneumoconiosis	Pulmonary Embolus
Congestive Heart Failure	Pulmonary Hypertension
Diffuse Idiopathic Pulm. Fibrosis	Pulmonary Neoplasm
Drug Related Pneumonitis	Pulmonary Tuberculosis
Emphysema	Rheumatoid Lung Disease
Goodpasture's Syndrome	Sarcoidosis
Gram Negative Pneumonia	Silicosis
Histiocytosis X	Spontaneous Pneumothorax
Hodgkin's Disease	Staphlococcal Pneumonia
Influenza	Traumatic Pneumothorax
Wegener's Granulomatosis	

preceding items to be evaluated (arith item) , or a probabilistic expression using sensitivity and specificity of an observation (prob item) (24).

After a search item has been created using the desired text string from PTXT, HCOM allows a set of five constraints to be defined, which will further define the data appropriate for the decision to be made. FROM can be used to create a time constraint for the search. TO allows an upper bound to be placed on each search. MOD (modifiers), such as first, last, maximum, minimum, can be applied to search items to further specify what information to retrieve. IF can be used to create conditional statements that are useful in selecting only certain fields that are to qualify. USE allows one to assign any value to the search item for later reference in a subsequent arithmetic item (24).

Arithmetic items may use any kind of mathematical functions needed to represent the logical model of a decision. An arithmetic item may be used to derive a value to be assigned to that item or any other item label for future reference, or to perform conditional termination or branching ahead in the sector (24).

Bayesean probability may be implemented directly with HCOM using the PROB item. There are four parameters in a PROB item: 1) the item whose value or existence is to be used, 2) the probability of the decision being true before consideration of the above item, 3) its sensitivity, and 4) its specificity in the context of the decision represented by this sector. The calculated a posteriori probability can be used as the a priori probability in a subsequent PROB item (24).

A set of final evaluations can be specified using arithmetic operators to manipulate any items in the sector. The values of a final evaluation statement can be inserted into a position occupied by an equal sign in the sector text, or as an index to select a sector text modifier to insert at that point in the text (24).

HCOM, within HELP sectors, has the ability to ask for data that are not found in the patient's file. This item specifies who is to be asked (patient, doctor, or nurse), whether the data are to be requested hierarchically, and which search items should be requested if the data are not already in the patient's file (24).

Figure 1 is an example of a HELP sector that determines a patient's likelihood of having bacterial pneumonia using the patient's historical findings. As the sectors are designed in a modular fashion, they are organized into "blocks" that deal in a particular medical area. This history sector begins by listing which block of sectors it belongs to. In this case the block is "PULMONARY SECTORS." The next line contains the title of the sector, "SECTOR 1 =.== PNEUMONIA (HISTORY)." The Final Evaluation is the a posteriori probability of the Bayesian constructs that the sector has calculated. It is

```
Block #7.141    PULMONARY SECTORS
Sector 1  =.== PNEUMONIA (HISTORY)

FINAL EVALUATIONS:
A       VAL: L
B       VAL: L / A

SECTOR LOGIC:
A       ARITH:                  0.0672
B       SEARCH:             ^   (A) [FC] CARDIO-PULMONARY, [N] HAVE YOU HAD
                                    RECENT CHEST PAIN?, FROM: 0, TO: NOW,
IF EX: VAL SUBITEM A GE 5
C       SEARCH:             ^   (A) [FC] CONSTITUTIONAL, [N] HAVE YOU HAD A FEVER
                                    WITH THIS ILLNESS?, (B) [N] HAVE YOU HAD CHILLS
                                    WITH THIS ILLNESS?, FROM: 0, TO: NOW,
IF EX: A OR B, USE VAL: MAX(A,B)
D       SEARCH:             ^   (A) [FC] CARDIO-PULMONARY, [N] HAVE YOU HAD A
                                    COUGH WITH THIS ILLNESS?, FROM: 0, TO: NOW,
IF EX: A
E       SEARCH:             #   (A) [FC] CARDIO-PULMONARY, [N] HAVE YOU HAD
                                    RECENT CHEST  PAIN?, [ADJ] IS YOUR CHEST PAIN
                                    INCREASED BY BREATHING DEEPLY?, (B) [ADJ] IS YOUR
                                    CHEST PAIN INCREASED BY COUGHING?, FROM: 0,
TO: NOW, IF EX: A OR B, USE VAL: MAX(A,B)
F       SEARCH:                 (A) [FC] CARDIO-PULMONARY, [N] HAVE YOU BEEN
                                    SHORT OF BREATH WITH THIS ILLNESS?, FROM: 0,
        TO: NOW, IF EX: A
G       SEARCH:             #   (A) [FC] CARDIO-PULMONARY, [N] HAVE YOU HAD A
                                    COUGH WITH THIS ILLNESS?, [ADJ] DOES YOUR COUGH
                                    BRING UP ANYTHING?, [ADV] IS THE SPUTUM YELLOW,
                                      GREEN OR BROWN?, FROM: 0, TO: NOW, IF EX: A
H       PROB:                       A,  IF EX: C,  USE VAL: C,  MIN: (1, 5),
TRUE: (0.42, 0.58),  FALSE: (0.832, 0.168)
I       PROB:                       H,  IF EX: D,  USE VAL: D,  MIN: (1, 5),
TRUE: (0.06, 0.94),  FALSE: (0.494, 0.506)
J       PROB:                       I,  IF EX: E,  USE VAL: E,  MIN: (1, 5),
TRUE: (0.71, 0.29),  FALSE: (0.85, 0.15)
K       PROB:                       J,  IF EX: F,  USE VAL: F,  MIN: (1, 5),
TRUE: (0.22, 0.78),  FALSE: (0.45, 0.55)
L       PROB:                       K,  IF EX: G,  USE VAL: G,  MIN: (1, 5),
TRUE: (0.58, 0.42),  FALSE: (0.82, 0.18)
M       ARITH:                      IF L LT A THE GOTO FE
N       EXIST:                      ASK((PATIENT QUESTIONS) C, D, E, F, G)
```

Figure 1. Example of a diagnostic HELP sector for pneumonia. Parts of the sector include: 1) sector label, 2) final evaluation slot, 3) a priori probability for this disease, 4) search items indicate the questions required to calculate disease likelihood, 5) prob items specify the statistics (sensitivity and specificity) associated with yes and no answers to referenced questions, 6) control logic for ASK function, 7) specification of questions to ask patient.

returned and stored in the patient database as the final probability for the particular sector that has been run on the patient's database.

The next section of the sector consists of the "SECTOR LOGIC." The first statement is an arithmetic statement that assigns the a priori probability, the probability of the disease occurring in the general population, of 0.0672 to item A. Items B through G are search items. These items cause the patient's database to be searched for the answer the patient gave to these questions. They are responsible for determining the presence or absence of these historical findings in the patient's database.

Both modes of data collection used in this research stored the questions asked with the corresponding answers on the patient's database. A "yes" answer is stored numerically as a 5, a "no" answer was assigned the value of 1, and a question not asked was equal to 0. These values come into play in the probability section of the HELP sector to determine which statistics to use, but are also used in the search items. For example both item C and E are searching for two questions. The "USE VAL: MAX(A, B)" assigns the maximum value of the two questions found to the item. So that if one of the questions was answered yes [5], and the other was answered no [1], then the value of 5 would be assigned to this item for use in the probability section.

Items H through L are the probability section of the sector. Here each answer is processed using a Bayesian operation that calculates the likelihood of the disease based upon the presence or absence of a specific symptom (answer to the question). The first terms (A,H,I,J,k) are the a priori probabilities for each corresponding Bayesian item. The a posteriori probability of each "PROB" item becomes the a priori probability of the next "PROB" item. This method is a sequential application of Bayes' Theorem (25).

In this example the "PROB" item checks to see if a "SEARCH" item exists. If it does, it associates the answer (no = 1, yes = 5) with the correct conditional probabilities (true positive, false positive, true negative, false negative). For example, if the patient answered "yes" to the question "Have you been short of breath with this illness?", the Bayes function would assign 0.78 as the likelihood that a patient with pneumonia will be short of breath, and 0.55 as the likelihood that a patient that does not have pneumonia will be short of breath. The Bayes function uses these values to calculate the probability of pneumonia in this patient complaining of shortness of breath. This is done for each "PROB" item beginning with item H and continuing sequentially through to item L, where a final a posteriori probability of this patient having pneumonia is obtained. This is then stored in the patient database via the final evaluation statement. Figure 2 is an example of how the Bayes function would work for the Bacterial Pneumonia sector listed in Figure 1.

FINDINGS

ITEM	ANSWER	TEXT
B	NO = 1	Have you had recent chest pain?
C	YES = 5	Have you had a fever with this illness?
	NO = 1	Have you had chills with this illness?
D	YES = 5	Have you had a cough with this illness?
E	NO = 1	Is your chest pain increased by breathing deeply?
	NO = 1	Is your chest pain increased by coughing?
F	YES = 5	Have you been short of breath with this illness?
G	YES = 5	Does your cough bring up anything?
	YES = 5	Is the sputum yellow, green or brown?

BAYES' THEOREM

$$P(D/S) = \frac{P(D) * P(S/D)}{P(D) * P(S/D) + P(ND) * P(S/ND)}$$

P(D/S) = probability of disease D given symptom S
P(D) = the a priori probability of disease D
P(S/D) = the sensitivity of symptom S for disease D
P(ND) = probability of not having disease D
P(S/ND) = the false positive rate (1 - specificity) for symptom S in disease D

ITEM	A PRIORI	FINDING	VALUE	S/D	S/ND	A POSTERIORI
H	0.0672	C	YES	0.58	0.168	0.1992
I	0.1992	D	YES	0.94	0.506	0.3161
J	0.3161	E	NO	0.71	0.85	0.2785
K	0.2785	F	YES	0.78	0.55	0.3538
L	0.3538	G	YES	0.42	0.18	0.5609

Probability, after running history frame, of having bacterial pneumonia = 56.09%

Figure 2. Example of the sequential application of Bayes equation used for the set of historical findings found in the Bacterial Pneumonia sector listed in Figure 1.

The sector in Figure 1 represents only one half of the logic used for determining bacterial pneumonia, that of historical finding. The second half, or clinical data sectors, contains a list of clinical data findings specific to bacterial pneumonia. It deals with physical examination, laboratory, and radiology findings. There are generally two sectors for each disease. The a posteriori probabilities generated by the history sectors are used as the entrant a priori probabilities of the corresponding clinical data sectors.

Data Dictionary: PTXT

In order for the HELP system to be able to use the information stored on the clinical database, in this case the answers to the computerized patient history, the data have been stored in a uniquely coded form. Thus coded (as opposed to free text), the data can be easily retrieved and analyzed for use in research and decision logic. The code for each term is derived from the position of the defined term in a hierarchical structure representing medical terminology. A record exists for each medical term that is defined in the data dictionary. There are several fields in the record. The first is the hierarchial code that acts as a unique key for the entity. The medical terminology (text) is stored in the second field. The dictionary is called PTXT (pointer to text) because of this fundamental relationship. Keywords are also associated with each entry. Keywords are used to reference data elements already defined in the data dictionary (24).

The hierarchial structure of the PTXT dictionary is designed to optimize the speed with which data items can be stored or retrieved (22). The hierarchy consists of the following levels (from top down): data class, data type, field code, noun, adjective, adverb. Modifying terms can also be defined to apply to multiple terms within a hierarchy (24). The first level, data class, is used to organize data into related areas of medicine. Different data sources within the hospital are assigned to separate specific data classes. It defines which general data class the data item is associated with. For example, Data Class 20 is used to define Radiology findings while Data Class 7 is used for history and physical exam findings.

Data type indicates the format in which the data are stored in the actual record in the computer. Type zero is the simplest and most efficient way to store a fixed-length string. This is especially useful when all the data in a field are ready to be recorded together each time any one of them is recorded. The clinical laboratory uses this format when recording a SMAC 20 result, in which every word in the string represents a particular chemistry value whose units are defined in the PTXT definition (22).

Type one data are organized into a further hierarchy beyond the data class and field code, with the content of the string being self defining. Delimiters precede type one data

codes that define what kind of item will occur next in the string. This format is very flexible and is useful when constructing data from questionnaires in which the number and types of answers vary among patients (24).

Type three is used to store decisions made by the HELP system from other data stored in the patient's file. It gives the data class, field code, HELP sector number, value of the sector if appropriate, and any sector modifiers. All of the alerts, data interpretations, treatment suggestions, and diagnoses that have been made by the system are stored as type three data (24).

Field codes are the next level below data type in the PTXT hierarchy. This is the initial division of the data class into smaller related areas. For instance, in Data Class 7, Field Code 120 represents a block of questions for cardio-pulmonary history.

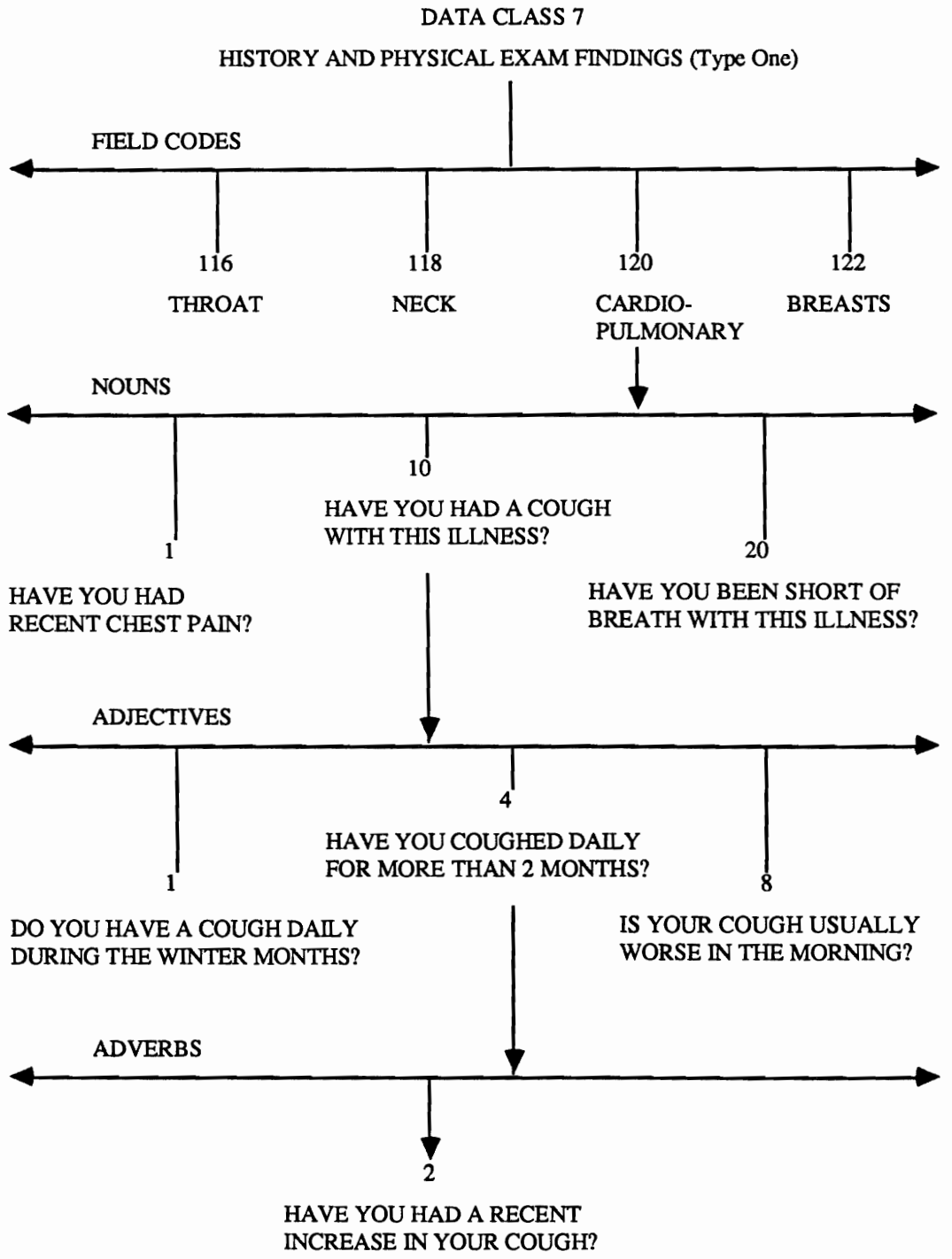
The next levels of classification are represented by nouns, adjectives and adverbs. Nouns define specific items within the field code. Adjectives are associated with and qualify each noun, and adverbs are associated with and qualify an adjective.

For example, in Field Code 120, Noun 10 is the question "Have you had a cough with this illness?" Adjective 4 associated with Noun 10 is "Have you coughed daily for more than 2 months?" Adverb 2 qualifying Adjective 4 is " Have you had a recent increase in your cough?" Figure 3 is an example of how the PTXT hierarchical data base structure is arranged.

GQAP

General Question Asking Program (GQAP) is a data entry subsystem that facilitates the acquisition of data entered from a terminal keyboard. QSTN, a screen editor program within the GQAP system, makes it possible for technical personnel without programming skills to construct questionnaires. The created questionnaire links the entered data to PTXT, provides error checking of the entered data, and provides for user defined logic for control of the screen presentation. GQAP executes the questionnaire and stores the entered data in the patient database (24).

The following seven basic screen types are available to the user when constructing a questionnaire: 1) multiple choice entry from a menu, 2) value (data) entry where a list of items are presented for entry of numeric values, 3) keyword entry where a keyword is entered to access the data dictionary to locate the exact item for entry into the patient's database, 4) entry of cost accounting codes 5) entry of free text data, 6) entry of time and data, and 7) no user entry question for entry of constant or fixed strings into the data base. The screens can be sequenced by a fixed presentation order or a branching order contained in the definition of the screens (21).



PTXT code = 7-1-120-10-4-2

Figure 3. Example of the hierarchical data base structure of PTXT, showing the PTXT code for the question "Have you had a recent increase in your cough?"

When defining data entry and multiple choice screens it becomes necessary to also define the header, the fields, the diagnostic logic, and the follow-up logic. The header or prompt message is displayed at the top of the screen with its primary function being to instruct the user on the entry of data for that screen (24).

The developer must determine, for each field on the screen, the text which is to be displayed on the terminal, the associated PTXT codes and any special characteristics of the field. Diagnostic logic, which can be used to perform limited checks on entered data, can also be included with each screen. These diagnostics check for ranges of numeric entries, consistencies of multiple choice selections, inclusion of numeric values, allowable time entries, etc. (24).

As many follow-up statements as necessary may be included in the definition of the screen. If more than one statement is included and satisfied then both follow-up screens will eventually be presented to the user. The general format of a follow-up statement is "ASK (question number) if (Boolean statement)." In developing a questionnaire, an initial question or set of questions is declared as key questions and is immediately placed on the "screens to be displayed stack" when the questionnaire is activated. The control of screen presented is then accomplished by "popping" the "screens to be displayed stack" and presenting the screen at the top of the stack. The follow-up logic causes more screens to be added to the stack and the presentation of a screen causes it to be removed from the stack. When the stack is empty, the questionnaire is terminated and control returned to the HELP system for initiation of other applications (24).

GQAP Computerized Patient History Collection

One of the computerized data gathering techniques used and compared in this research project was a questionnaire that was created using QSTN, and then executed by GQAP as described above. The questionnaire was created using the patient history questions found in the 41 pulmonary and cardiology HELP sectors listed in Table 1. This was done to make certain every history question, which those HELP sectors searched for in the patient database when they were run, would be stored in the patient database by GQAP. After GQAP had finished presenting the questionnaire and storing the data on the patient database, the 41 pulmonary and cardiology HELP sectors were run on that patient's database. With the sectors run and the respective probabilities stored for each disease, a differential diagnostic list of the top five diseases was then generated for each patient. This was used as a means of comparison of accuracy between GQAP data collection and the DDA (decision-driven data acquisition) mode of history collection, which will be explained later in detail.

The questionnaire that was created consisted of a total of 127 screens. The screen types used in this questionnaire were data entry and no user entry. The data entry screens were the ones that were presented to the patients as the questionnaire gathered their history. These screens began with a prompt message that instructed the patient as to how the questions on that screen were to be answered. For example, most of the screens began with the prompt: "Press "Y" for yes or "N" for no for each of the following questions."

As GQAP executes the questionnaire, the screen that is presented to the patient consists of the prompt followed by a list of questions, usually a maximum of five, to which the computer expects an answer in the form of a value to be given. Each question is then listed individually so that the patient can answer each question, one at a time. Figure 4 is an example of what a screen presented to the patient by GQAP would look like. When the questionnaire was created these questions were given the value of zero. As the questions are answered their value is changed to either one or five. On the terminal that was used for the purpose of gathering these histories, we placed a "Y" over a key that entered a numerical 5 as an answer, and a "N" over a key that entered a numerical 1 as the answer to the question. Therefore as the patient answers each question either "yes" or "no", a value of either 5 or 1 respectively is then stored for each question.

The no user screens were implemented for answering "no", giving a value of 1, to questions which, in PTXT, were hierarchically beneath questions to which the patient had already answered "no" to. They not only save the patient from having to answer obviously repetitive "no" answers, but also saved time in the history collecting process. These screens were not presented to the patient, but were answered automatically by GQAP so that the patient would not have to waste the time it would have taken to answer them themselves. For example, if the patient had answered no to the question "Have you had wheezing with this illness?", in the follow up logic to that question the questionnaire would have branched to a no user question that would have assigned the value of 1 to the following questions: 1) "Do you wheeze due to an allergy?" 2) "Do you wheeze due to an infection in your lungs?"

The questionnaire was created to be branching in order. The branching that occurs is defined in each screen through the follow-up logic. There is just one key question, question number 1, that is placed immediately on the "screens to be displayed stack." The follow-up logic for each question causes more questions to be added to the stack. Each screen not only causes the history questions that, in PTXT, are hierarchically beneath those in the present screen to be asked, but also causes the next question in line to be put on the "screens to be displayed stack." Examples of screens from the patient history questionnaire are found in Figures 5 through 7.

PRESS "Y" FOR YES OR "N" FOR NO FOR EACH OF THE FOLLOWING QUESTIONS:

1..HAVE YOU HAD A FEVER WITH THIS ILLNESS?
2..HAVE YOU BEEN SWEATING MORE THAN USUAL?
3..HAVE YOU HAD CHILLS WITH THIS ILLNESS?
4..HAVE YOU BEEN CONFINED TO BED WITHIN THE LAST TWO WEEKS?
5..HAVE YOU BEEN LOSING WEIGHT?

1..HAVE YOU HAD A FEVER WITH THIS ILLNESS?__

Figure 4. Example of user entry screen, in the GQAP history collection program, as seen by the patient.

QUESTIONNAIRE NUMBER: 163
QUESTION NUMBER: 7
QUESTION TYPE: DATA ENTRY
MIN RESPONSES: 0
MAX RESPONSES: 0

ERASE SCREEN BEFORE QUESTION: YES
TYPE TO BE PACKED: DATA
STORE STRINGS AFTER QUESTION: YES
SAVE TEXT FOR REVIEW: NO

PRESS "Y" FOR YES OR "N" FOR NO FOR EACH OF THE FOLLOWING QUESTIONS:

	PTXT CODE	
1...HAVE YOU HAD A FEVER WITH THIS ILLNESS?	7 1 100 1 2 0 0 0	VALUE:0
2...HAVE YOU BEEN SWEATING MORE THAN USUAL?	7 1 100 1 6 0 0 0	VALUE:0
3...HAVE YOU HAD CHILLS WITH THIS ILLNESS?	7 1 100 1 10 0 0 0	VALUE:0
4...HAVE YOU BEEN CONFINED TO BED WITHIN THE LAST TWO WEEKS?	7 1 100 1 12 0 0 0	VALUE:0
5...HAVE YOU BEEN LOSING WEIGHT?	7 1 100 1 14 0 0 0	VALUE:0

FOLLOW UP LOGIC:

1...ASK 68 IF 5=(C5)
2...ASK 163 IF 5=(C1)
3...ASK 67 IF 2=(C5)
4...ASK 168 IF 2=(C1)
5...ASK 8 IF (C1)

DIAGNOSTIC LOGIC:

NO DIAGNOSTICS ESTABLISHED

QUESTION 0007 CALLED BY THE FOLLOWING QUESTIONS:
6

Figure 5. Example of question number 7 from QSTN questionnaire.

```
QUESTIONNAIRE NUMBER: 163
QUESTION NUMBER: 68
QUESTION TYPE: DATA ENTRY
MIN RESPONSES: 0
MAX RESPONSES: 0

ERASE SCREEN BEFORE QUESTION: YES
PACK AFTER QUESTION: YES
TYPE TO BE PACKED: DATA
SAVE TEXT FOR REVIEW: NO

PRESS "Y" FOR YES OR "N" FOR NO FOR EACH OF THE FOLLOWING QUESTIONS:

                                                    PTXT CODE
1...HAVE YOU LOST 10 POUNDS OR MORE DURING THIS ILLNESS?
                                                    1 100 2 14  2 0 0       VALUE:0
2...HAS YOUR WEIGHT LOSS BEEN OVER MONTHS OR YEARS?
                                                    1 100 2 14  4 0 0       VALUE:0
3...HAVE YOU HAD A RECENT WEIGHT LOSS?              7 1 100 2 14  6 0 0     VALUE:0

FOLLOW UP LOGIC:

NO FOLLOW UPS ESTABLISHED

DIAGNOSTIC LOGIC:

NO DIAGNOSTICS ESTABLISHED

QUESTION 0068 CALLED BY THE FOLLOWING QUESTIONS:
7
```

Figure 6. Example of QSTN format of question number 68, which is a follow-up logic question for question number 7.

QUESTIONNAIRE NUMBER: 163
QUESTION NUMBER: 163
QUESTION TYPE: NO USER ENTRY

ERASE SCREEN BEFORE QUESTION: YES
PACK AFTER QUESTION: YES
TYPE TO BE PACKED: DATA
SAVE TEXT FOR REVIEW: NO

1...HAVE YOU LOST 10 POUNDS OR MORE DURING THIS ILLNESS?
1 100 2 14 2 0 0 VALUE:1
2...HAS YOUR WEIGHT LOSS BEEN OVER MONTHS OR YEARS?
1 100 2 14 4 0 0 VALUE:1
3...HAVE YOU HAD A RECENT WEIGHT LOSS? 7 1 100 2 14 6 0 0 VALUE:1

FOLLOW UP LOGIC:

NO FOLLOW UPS ESTABLISHED

DIAGNOSTIC LOGIC:

NO DIAGNOSTICS ESTABLISHED

QUESTION 0163 CALLED BY THE FOLLOWING QUESTIONS:
7

Figure 7. Example of a No User Entry question. Question number 163 in QSTN format, which is a follow-up logic question for question number 7.

Decision-Driven Data Acquisition (DDA)

DDA is the other computerized data collection method that was implemented for comparison in this project. DDA is a recent addition to the HELP system and represents an effort to bring built-in system intelligence to the problem of qualitative data collection (21). The goal of this system is to ask the questions that are best suited to elaborate the most likely diagnoses (26). This frame-based decision system is used to direct a computer-administered history and to generate a 1 to 5 member differential diagnostic list based only upon this history. The frames used in the decision system, for this project, are the cardio-pulmonary sectors listed in Table 1. A Bayesian scoring algorithm allows this system to recognize the most likely diseases and to choose questions to ask that will be useful in elaborating these diagnoses (27).

DDA uses the concept of hypothesis directed questioning in a cyclic process of hypothesis generation followed by the collection of data necessary to explore these hypotheses (27). These hypotheses are system decisions represented by HELP sectors. As the computer begins the DDA system of history data acquisition an initial set of key questions are presented to the patient (27, 28). The questions are answered by the patient, who presses one of two special keys on the terminal keyboard marked "YES" and "NO." The six key questions in the pulmonary history gathering system are:

1. Have you had a fever with this illness?
2. Have you ever had asthma?
3. Have you been short of breath with this illness?
4. Have you had a cough with this illness?
5. Have you had recent chest pain?
6. Have you had wheezing with this illness?

When these questions are answered they are used to evaluate the diagnostic frames in the knowledge base. For example, if the patient answers a key question positively, the stored question and response will data-drive HELP sectors that use that question as a data-driving item. As the system evaluates these sectors a new set of questions is selected and submitted to the patient. The new questions are derived from from ASK items that exist in the newly evaluated sectors. These questions represent the systems assessment of the most useful information to request. This cycle of asking new questions that are then used to reevaluate the diagnostic frames in the knowledge base to calculate new probabilities and generate new questions, continues until the patient has been queried for all the data relevant to their current pulmonary health problem and/or a diagnosis of significant probability is obtained (26, 27, 28, 29).

Two tools, which were referred to briefly above, are built into the HELP system to manage this cycle. The first process is designed to provide a method for new data to trigger the evaluation of a diagnostic module (sector) whenever prespecified historical data is captured and stored. The frame author indicates which data used for the decision will evoke (data-drive) that decision by placing a flag (^) in front of the appropriate slots (26). The HELP sector for pneumonia listed in Figure 1 illustrates the use of data-drivers. In part (24) of the frame, the search items, the data-driving flag (^) has been placed in front of the following questions: "Have you had recent chest pain?" "Have you had a fever with this illness?" "Have you had chills with this illness?" and "Have you had a cough with this illness?". An answer to these questions will cause the logic in this frame for pneumonia to be processed. The initial set of diagnostic hypotheses is built on the answers to the set of key questions (listed above) after which all further questioning is driven by these and subsequent hypotheses (27).

The second tool, the ASK function, is designed to indicate which questions should, if their answers are not known, be asked of the patient. In Figure 1, for example, if the probability following the evaluation of the available information is greater than or equal to the a priori probability before the frame was run (item M), the ASK function (item N) will cause additional questions to be directed to the patient. The ASK function is designed to be "intelligent," in that it will only ask those questions which, after searching the patient's data base, are indeed absent (have not been answered before) (26). This function localizes the control of data collection within the sector and provides a truly hypothesis (decision)-driven questioning tool (28).

As the ASK function is activated, questions that are to be asked are sent to a buffer. At this point a separate process, the QUERY DRIVER, is responsible for formatting the data requests stored by the ASK function and for presenting them to be asked. After gathering the questions sent to a special buffer by the ASK function, it determines which five should be directed to the patient. The patient's answers to the new set of questions are stored and the frames that sent them are reevaluated to determine the effect of the new information on the likelihood of those diseases. New frames may be also be triggered for which these answers are data-drivers. The new frames may in turn contribute questions of their own to the ASK buffer, which may be directed to the patient in the next questioning cycle (26, 27, 28).

In some cases, the QUERY DRIVER can take advantage of the hierarchical structure of the data base to infer answers to questions by referring to previously answered questions. The questions used in the DDA program use the same hierarchical data dictionary described above in the PTXT section, with a general upper level question

branching to lower more specific questions. A "no" answer to an upper level question implies a negative response to those questions hierarchically below it and they will not be asked, but a "yes" answer generates a list of hypotheses for which answers to the more specific, second level questions may be necessary. The use of this hierarchical data base also makes it necessary for the QUERY DRIVER to ask and store parent questions before it is able to ask lower level questions. This data structure reduces the questioning process and also eliminates the confusion the patients might feel if during the questioning process they were asked detailed information about a general symptom that they did not have (26, 28).

Two other control mechanisms are built into the QUERY DRIVER that help manage the flow and direction of the questioning process. The first is a question selection algorithm used to decide which of the questions in the buffer it will ask. It is based upon the assumption that the most satisfactory history will be collected by attempting to match the data requirements of the most likely diseases. The QUERY program makes the selection by summing for each question in the question buffer the likelihood of the disease frames that sent that question to the buffer. Therefore, questions whose answers would contribute to more than one diagnosis tend to score higher than questions that are used by a single diagnostic hypothesis. The totals for each question are then compared and the process selects the top five questions to present to the patient (26, 27, 28).

The second control mechanism is a history termination algorithm that is designed to end questioning when it appears unlikely that further questioning will significantly alter the probabilities of any of the hypothesis under consideration. This process is invoked after the QUERY program asks the first 30 questions, and again after each group of 5 questions. The algorithm sums the probabilities of the unexplored diseases that have questions remaining in the question buffer. If the sum of the probabilities of these incompletely explored hypotheses is less than 0.05, the QUERY program terminates the questioning process (26, 27).

Method of History Collection and Patient Population Used

This project focused on the entry of data into a computer-based medical information system by the patients themselves. We used a terminal mounted near the bedside, or a mobile terminal that could be rolled from room to room in order to collect a history of present illness directly from the patient.

The subjects for the project were a group of patients with pulmonary and/or cardio diseases. They were patients of a select group of local specialists and subspecialists who

had agreed to let their patients participate, if they (the patients) agreed to do so. These physicians also agreed to help evaluate the relative value of the two data entry tools (21).

History collection using the DDA-QUERY program was begun in January 1985 and continued through January 1986. The QSTN-GQAP history collection mode began in February 1986 and continued until August 1986. During this time 52 DDA-QUERY histories and 24 QSTN-GQAP histories were collected.

The basic procedure for obtaining a history from a patient was essentially the same using either mode of data collection. One member of a group of research personnel would check the hospital information system to ascertain whether or not a patient of one of the participating physicians had been admitted to the hospital. Only patients who had been admitted to the hospital for no more than two days were used for this research. It was observed that patients who had been admitted longer than two days either forgot their original symptoms or began to complain of side effects caused by their being in the hospital, such as effects from treatment procedures, operations, or hospital acquired diseases. Their room number was noted and if it were not one in which a bedside terminal existed a rolling terminal was taken to the patient's room.

The patients were informed that taking their history with the computer was a research project and that their physicians had consented to let them participate if they would. They were also informed to enter their histories as they applied to the disease(s) they had entered the hospital with. The researcher then hooked up the rolling terminal through the phone line to the main computer system and showed the patients how to enter their histories using the computer.

Help was given to the patient as it was needed. Sometimes questions had to be read or help with entering the data was needed. Most patients were able to use the program on their own once it was started and explained to them.

After the history had been gathered, a report generating program was run that listed the history of present illness and gave a differential diagnostic list of what the computer had decided the top five diseases of the patient were. These were then put on the patient's chart and a copy was sent to the attending physician. Input was requested from the physicians as to how they felt about the accuracy of the history collection mode and for any suggestions they might have.

Methods of Comparison and Evaluation

This experiment was set up with the following null hypothesis: There is no measurable difference in 1) time to complete, 2) number of questions asked, or 3) accuracy

of decisions made from the entered data using a conditional branching set of fixed frames (QGAP) as opposed to a decision-driven data acquisition (DDA) system (21).

As each questioning process was begun a time-stamp was entered automatically into the patient's record. Another time-stamp was entered when the questioning process had been completed. This served to measure the duration of terminal interaction the patient experienced during the history collection process. This duration time can used as an indirect measure of ease of entry for both questioning procedures, and more importantly could be a determinant of the patient's responses to this data collection experience. The longer the duration, the more anxious the patient is to have the whole process over with, especially the more ill the patient is. The duration time was also related to the type and quality of data entered and the decisions, driven or evoked, by these data for each user.

During either history collection process, as the patient answers questions, those answers are stored in the patient's clinical data file. With all questions and responses stored, we were able to determine how many total questions were answered in each session, how many yes and no questions the patients themselves had entered, and how many questions each program had answered itself without the patient actually inputting answers. Again, as above, the difference in the amount of questions asked by each history collection mode (GQAP vs. DDA-QUERY) was related to the type and quality of data entered and the decisions, driven or evoked, by these data for each user.

In order to determine the accuracy of the decisions made using the responses obtained in each history collection session, we were able to take advantage of the fact that at the time of entry an admitting diagnosis is entered of each patient. Also, after the patient has been released from the hospital, the medical records department (using ICD9 codes) enters the "true" discharge diagnoses that were confirmed during the patient's hospital stay. Both the admitting and the discharge diagnoses were compared to the five member differential diagnostic list generated using the entered data obtained from the history collection programs.

A program was written that would search each patient file and return a condensed report. Contained in this report were the essential parameters needed to compare and evaluate each collection mode. It listed patient demographics (name, admit date, etc.), the admitting diagnosis, history start time, history stop time, duration of terminal interaction, questions answered yes, questions answered no, number of questions asked in the session, the number of diseases considered, the differential diagnoses (with scores), and the discharge diagnoses. The information collected in this program was used as the source for comparison between the two different history collection modes.

Methods for Updating X-ray Statistics for Use in Sector Logic

A physical exam from the physician was the next type of data that was scheduled to be collected at the patient's bedside (21). When the diagnostic frames were first developed, there were 58 frames for 29 diseases. There were two frames for each disease. The first dealt with patient history (as described above) and the second dealt with physical exam, laboratory, and radiology findings. The history frames (sectors) have evolved to include more than the original 29 disease states (Table 1), whereas the physical exam frames have continued to include only the original 29 diseases (Table 2). Before collection of a physical exam could begin on a regular basis inconjuction with the collection of a patient history, the physical exam frames needed to be updated.

When the diagnostic frames were originally designed, a group of medical experts were assembled to select diseases for inclusion in the system , to indicate which clinical data were most useful for diagnosing each disease, and to estimate a set of numeric parameters for each disease including an a priori probability for that particular disease, and a sensitivity and specificity for each manifestation in the disease frame (23). The a priori probability, sensitivities and specificities for the history frames had recently been revised during an earlier study, using a patient data base collected at LDS Hospital (30). The same needed to be done for the physical exam frames.

This section deals with revising the statistics of radiology findings used in the physical exam diagnostic frames of the pulmonary diseases listed in Table 2. The goal in this phase of the project was to get the physical exam frames up and running with the most accurate data available at this time. It was felt that revising the statistics for the radiology findings, using a data base of actual patients who had entered LDS Hospital, would create a better set of statistics than the ones that were originally estimated by the medical experts. This had been the case with the history findings, which had been revised in a similar fashion in an earlier study (30).

To test the effects of the revisions and to ascertain that they were indeed improvements over the original estimates, both the old set of statistics and the new set were run against a training and a test set of patients. The respective differential diagnostic lists that were generated using the different statistics (estimated vs. calculated) were compared to the known diseases of the patient's to observe if there was a difference between the sets of statistics and if in fact an improvement had been made.

This section includes a description of the patient data base used for the purpose of revising the statistics for the radiology findings including the training and test sets used, the subsystems within the HELP System which were used during the revision and testing

Table 2. The 29 diseases included in physical exam frames.

Acute Bronchitis	Histiocytosis X
Asbestosis	Hodgkin's Disease
Aspiration Pneumonia	Influenza
Asthma	Lung Abscess
Bacterial Pneumonia	Metastatic Neoplasm
Bronchiectasis	Non-Hodgkin's Lymphoma
Chronic Bronchitis	Primary Pulmonary Neoplasm
Coal Worker's Pneumoconiosis	Primary Pulmonary Hypertension
Coccidioidomycosis	Pulmonary Embolism
Congestive Heart Failure	Sarcoidosis
Diffuse Idiopathic Fibrosis	Silicosis
Drug Related Pneumonitis	Spontaneous Pneumothorax
Emphysema	Tuberculosis
Goodpasture's Syndrome	Wegner's Granulomatosis
No Pulmonary Disease	

processes (STRATO), how the statistics were revised, and finally how the revisions in the statistics for the radiology findings were evaluated.

Patient Data Base

The patient data base used for this study was the same one that had been created and used for the project dealing with the revision of the statistics for the history frames (30). The patients in the data base were all hospital in-patients who had been admitted a maximum of 72 hours and who had received a chest X-ray within 48 hours of admission. The patients in the data base were all patients of physicians who had given permission stating they could be included in the study.

Data were collected for a group of 637 patients entering the LDS Hospital. In addition to the information that is captured routinely by the HELP system, these patients had a patient history, physical assessment data and reported chest radiograph findings entered into their file. The patient history was gathered using a paper questionnaire which was later entered into the computer, and an interactive questionnaire was developed to collect a descriptive report of each patient's initial chest X-ray. This is all data needed by the decision frames to determine the patient's differential diagnostic list of pulmonary diseases.

In order to revise statistics for a particular disease, the patients in the data base needed to be classified into correct disease groups. Patients were initially assigned to a disease category on the basis of their discharge diagnoses, which are stored in the clinical data base using ICD-9 codes. These diagnoses are selected by the patient's attending physicians at the time of discharge and are entered into the computer by the medical records department. The primary discharge diagnosis should reflect the problem for which the patient was admitted to the hospital (31).

A table of ICD-9 codes that matched the 28 pulmonary diseases (Table 2) was compiled and used to classify the patients into disease categories. Patients without codes for any of the pulmonary diseases in that list were assigned into the No Pulmonary Disease group. An extensive review of the patient's charts was done to make certain ICD-9 codes stored by the medical records department were indeed an accurate reflection of the patient's "true" admission diagnosis. After classifying the patients into disease groups using their ICD-9 codes and checking that classification by way of a chart review, it was felt that the best possible "true" diagnosis classification had been obtained (30).

This group of 637 patients was divided into a training set consisting of 535 patients and a test set consisting of 102 patients. The training set was utilized for the revision of the statistics that the radiology findings access in the physical exam frames of the pulmonary

diseases. The test set was reserved to evaluate any changes in the diagnostic accuracy associated with the use of the new versions of these frames.

STRATO - The Research Subsystem

STRATO is a subsystem designed to support medical research on the patient data base being built with the HELP system. The name STRATO is derived from the first step in data analysis, the stratification of the larger patient data base into subgroups for data analysis. STRATO's main purpose is to detect in a clinical data base those patients matching given research criteria and to extract the specific information requested by the researcher (24).

There are four methods that STRATO can use to create the initial patient list. The first involves searching the clinical data base for patients with specific identifying or demographic data such as patient number, name, room, admit and discharge dates, etc. This type of search uses information in the patient's ID file. In the second method, criteria for selection are based upon the clinical data stored in the patient record. Searches can utilize laboratory results, radiology findings, medications taken, or a large variety of other criteria, in order to select a patient population (24).

Both of the above selection methods use an interactive, menu-driven command language present in STRATO, to construct search criteria. During the construction of search criteria, a process named ENTR mediates access to the HELP data dictionary. It uses key words to locate entries in the data dictionary that represent the required data. After the search data have been chosen, mathematical and logical combinations of these data can be specified and if required, further restrictions based on time or other modifiers can be entered. At last, the patient population upon which the search will be performed is chosen. This population can be a complete patient data base containing up to six months of inpatient and outpatient admissions, or a subset of such a data base that has been defined in a previous search (24).

The third method for patient selection involves the use of HELP decisions frames to identify search criteria. A frame is written in the HELP decision language that defines the characteristics of the desired patients. This is similar to writing a small program to define the search patients, and is best used for complex searches. The final selection method utilizes options within the STRATO program that define simple combinations of previously created patient groups. New patient groups can be produced from established groups by using the set operations of union, intersection, and formation of the compliment of the intersection (24).

A search results in the creation of a "population", which is a list of those patients who matched the criteria specified in the search. By doing a series of additional searches, the main population can be divided into subpopulations separated through the use of new search criteria.

STRATO's main functions are generating populations of patients who meet defined criteria, and extracting data from their records for later analysis. STRATO does not provide procedures for sophisticated statistical testing. Instead, there are programs available within the HELP system that accept STRATO files for further statistical analysis (24).

Revision of X-ray Statistics

To begin revision of the X-ray statistics, all 535 patients in the training set were entered into a STRATO file. A search was then performed upon this original set to obtain a population of patients who had chest X-ray findings stored in their files. Of the original 535 patients, 527 were found to have chest X-ray findings actually stored on file.

The patients in this data base had previously been classified into "true" disease groups, reflective of their actual admit diagnosis, as described above. These groups were entered manually into STRATO files. Of the patients in this training set, there were 28 with bacterial pneumonia, 3 with bronchiectasis, 15 with pneumothorax, 11 with pulmonary embolism, 3 with acute bacterial bronchitis, 14 with pulmonary metastases, 6 with aspiration pneumonia, 33 with chronic bronchitis, 12 with emphysema, 15 with asthma, 15 with pulmonary neoplasm, and 64 with conjestive heart failure. The rest were classified as having no pulmonary disease.

The training set was then analyzed to determine sensitivities (the probability of symptom given that the patient has the disease) and specificities (the probability of symptom given that the patient does not have the disease) for the radiology findings searched for in each particular disease frame. If an adequate number of patients had had a disease, it was possible to generate both sensitivities and specificities for the radiology findings which that disease frame referenced. It was decided that a minimum of six patients having a given disease was required before sensitivities could be derived from the data base. With specificities being based on the set of patients without a disease, new specificities could be derived for all of the disease frames.

During this stage of analysis STRATO was used to determine how many patients with or without a particular disease had the chest X-ray findings that were searched for within the disease frame that diagnosed that particular disease. For example, the physical exam sector for bacterial pneumonia searches for the following chest X-ray findings: lobar

consolidation or segmental consolidation or localized alveolar infiltrate. Therefore, in STRATO the population of patients with bacterial pneumonia is searched to extract those patients who have the chest X-ray findings of lobar consolidation or segmental consolidation or localized alveolar infiltrate. This STRATO search located 18 bacterial pneumonia patients who had the X-ray findings. As the bacterial pneumonia population consisted of 28 patients and 18 of those had the X-ray findings, the new calculated sensitivity for patients having the disease and also having the symptom is 0.64.

Similarly, STRATO was used to generate the specificities for the chest X-ray findings. To do this for the bacterial pneumonia frame STRATO is first used to create a population of patients by combining the population of the 527 patients having chest X-ray findings with the population of patients having bacterial pneumonia, using the set operation for the formation of the compliment of the intersection, to obtain a population of patients without bacterial pneumonia who have chest X-ray findings. This new population is searched by STRATO to extract those patients who have the chest X-ray findings of lobar consolidation or segmental consolidation or localized alveolar infiltrate. This search produced 27 patients without bacterial pneumonia who had the defined chest X-ray findings. Of the 499 patients without bacterial pneumonia who had X-ray findings, 27 had the desired X-ray findings; therefore the new calculated specificity for patients without the disease displaying the symptom is 0.054.

Sensitivities and specificities for the chest X-ray findings within the 29 physical exam disease frames were generated as described above. Ten of the 29 diseases were represented by sufficient patients so that both sensitivities and specificities could be generated for them. For the rest, only the specificities could be generated. Table 3 is an example of how the sensitivities and specificities changed for the chest X-ray findings used in the pulmonary neoplasm disease frame from the original estimated statistics to the revised statistics that were based upon an actual data base analysis.

After the training set had been analyzed and the new sensitivities and specificities had been generated, the pulmonary physical exam disease frames were revised. For the radiology findings represented in each frame the sensitivities and/or specificities were revised by replacing the old estimated statistics with the newly calculated ones, thus changing the true and false parameters used by Bayes in each disease frame. The a priori probabilities were also revised, by replacing them with ones that had been generated in the earlier study involving the revision of the pulmonary disease history sectors (30).

Table 3. Revision of statistics for Pulmonary Neoplasm based on data base analysis.

X-ray Finding	Original Sensitivity	Original Specificity	Revised Sensitivity	Revised Specificity
Solitary Nodule or Parenchymal Mass or Hilar Mass	0.59	0.95	0.33	0.979
Alveolar Infiltrate	0.34	0.85	0.13	0.88
Hypoaeration/Atalectasis	0.26	0.9	0.27	0.904
Hilar Adenopathy	0.3	0.8	0.13	0.988
Pleural Effusion	0.29	0.85	0.2	0.88
Multiple Nodules	0.03	0.95	0.13	0.9902
Bony Destruction/Erosion	0.1	0.95	0.02	0.99
Mediastinal Adenopathy or Mediastinal Mass	0.1	0.95	0.067	0.986

Evaluation of Revised Radiology Statistics

The next step, after revising the sensitivities and specificities for the chest X-ray findings in the pulmonary physical exam disease frames, was to determine whether or not an improvement had indeed been made over the original set of estimated statistics. This was done by evaluating their effect on the accuracy of the diagnostics system. The diagnostics results of the original frames were compared with those of the revised frames.

Both the training set and the test set were used in the evaluation of estimated vs. calculated statistics. Each set was tested in two ways. First the diagnostic frames were limited to access only the chest X-ray data available for each patient; second the frames were allowed to access both the history and chest X-ray data stored on each patient's file. In each case both the frames based on the original statistical estimates and similar frames using the derived statistics were evaluated against the same data set. This was accomplished by running all 29 frames in both groups (estimated and derived) against each patient in the training and test sets. A one to five member differential diagnostic list was then constructed that consisted of the most likely diagnoses but excluded any disease with a likelihood less than one percent.

There were four runs on each set (training and test) for which the computer generated a differential diagnostic list. The runs included the following combinations of data: 1) access of chest X-ray data only and use of estimated chest X-ray finding statistics, 2) access of chest X-ray data only and and use of derived chest X-ray finding statistics, 3) access to both history and chest X-ray data with the use of estimated chest X-ray finding statistics, and 4) access to both history and chest X-ray data with the use of derived chest X-ray finding statistics. The computer generated differential diagnostic lists created from each run were then compared to the final discharge diagnoses, the "true" diagnoses, stored on each patient's file. These lists were considered to be accurate when they contained a known discharge diagnosis (if none of the other diseases were present as discharge diagnoses then "No Pulmonary Disease" was considered appropriate) and inaccurate when a diagnosis was missed.

The computer generated differential diagnostic lists were evaluated according to the following features:

1). The rate for which the correct "true" diagnosis was included in the five member differential diagnostic list.

2). The rate for which the correct "true" disease ranked number one on the five member differential diagnostic list.

3). The rate for which the primary discharge diagnosis was included in the five member differential diagnostic list.

4). The rate for which the primary discharge diagnosis was ranked number one on the five member differential diagnostic list.

5). How many of the total pulmonary diseases which exist in each set (training and test) were captured in the differential diagnostic lists.

6). The rate for which the existing pulmonary diseases ranked number one on the five member differential diagnostic list.

The correct "true" diagnosis referred to above applies to those of the 29 pulmonary diseases in Table 2, pulmonary diseases and no pulmonary disease, which each patient was determined to have upon admission. It includes primary discharge diagnoses as well as secondary diagnoses. If a patient had more than one disease, each disease was evaluated individually (30).

McNemar's test was used to compare the accuracy of the differential lists produced by the original frames with those produced using the revised frames. Paired t-tests were used to compare the original and revised statistics.

RESULTS

This chapter contains the results of the comparison between the two modes of history collection (GQAP and DDA) and the results of the evaluation of the original estimated vs. the revised calculated statistics for the pulmonary findings used in the supporting data frames.

Results of the Comparison Between GQAP and DDA Used for Acquiring Patient Histories

History collection using the DDA-QUERY program began in January 1985 and continued through January 1986. The QSTN-GQAP history collection mode began in February 1986 and continued until August 1986. During this time 52 DDA-Query histories and 24 QSTN-GQAP histories were collected.

The disease distribution in both the QUERY and QSTN groups of patients was also compared to make certain that both groups of patients had roughly the same set of diseases. The researcher wanted to be sure that the diseases that the patients had were not biasing the results of the differential diagnostic lists that were produced based on the patients' answers to their computerized history questionnaires. The distribution of diseases in each group seemed to be quite similar. The patients in both groups were using the same basic set of logic frames in the data base to determine their differential diagnostic list of diseases.

The admit diagnoses for the patients in the QUERY group consisted of the following diseases: 3 had pneumonia, 1 had chronic bronchitis, 1 had congestive heart failure, 5 had emphysema, 10 had asthma, 1 had sleep apnea and 31 had other diseases that were not included in the Cardio-Pulmonary diseases listed in Table 1. The principle discharge diagnoses of this group of patients included the following diseases: 5 had pneumonia, 3 had chronic bronchitis, 3 had pulmonary embolus, 3 had emphysema, 10 had asthma, 1 had pulmonary neoplasm, 1 had sarcoidosis, 2 had acute bacterial bronchitis, 1 had metastatic neoplasm and 23 had other diseases not listed in Table 1.

The QSTN group of patients had admit diagnoses consisting of the following diseases: 6 had pneumonia, 2 had congestive heart failure, 4 had pulmonary embolus, 3

had emphysema, 4 had asthma, 1 had bronchiectasis, and 4 had other diseases not included in the list in Table 1. These patients had principle discharge diagnoses containing the following diseases: 5 had pneumonia, 1 had chronic bronchitis, 2 had congestive heart failure, 3 had emphysema, 4 had asthma, 2 had bronchiectasis, 1 had pulmonary tuberculosis, 1 had sarcoidosis, 1had metastatic neoplasm, and 5 had diseases other than the Cardio-Pulmonary diseases used in this research.

The chi-square test was employed to determine if there were statistically significant differences between these two groups (QSTN - Query admit diagnoses, and QSTN - Query discharge diagnoses). Because the number of patients in some of the disease categories was so small, the diseases were grouped into the following groups: infectious diseases (pneumonia, bronchiectasis, tuberculosis, bacterial bronchitis), obstructive diseases (emphysema, asthma, chronic bronchitis), neoplastic diseases (metastatic neoplasm, pulmonary neoplasm), pulmonary embolus, other pulmonary diseases (congestive heart failure, sarcoidosis, sleep apena), and nonpulmonary diseases. Since the logic sectors in the data base were concerned with the diagnosis of pulmonary diseases, the non-pulmonary disease category was not considered in this analysis.

In comparing the admit diagnoses between the two groups QSTN had 7 infectious diseases and Query had 3. QSTN had 7 obstructive diseases and Query had 16. Neither group contained neoplastic diseases. QSTN had 4 with pulmonary embolus while Query had none. Each group contained 2 with other pulmonary diseases. The chi-square test showed that the differences between these two admit groups were statistically significant with $p < 0.05$.

The comparison of the discharge diagnoses (considered to be the more accurate diagnosis of the two) showed QSTN having 8 infectious diseases and Query having 7. QSTN had 8 obstructive diseases and Query had 16. QSTN had 1 neoplastic disease and Query had 2. QSTN had none with pulmonary embolus while Query had 3. QSTN contained 3 with other pulmonary diseases and Query had 1 in this category. The chi-square test showed that there was no statistically significant difference between these two groups of discharge diagnoses.

Included in this section are the results of the comparison between the two history collection modes. The comparison between the two different modes was based upon the mean time it took to complete each respective history collection mode as well as the mean number of questions each mode required the patient to answer in order to complete a history collection session. The two modes were also compared based upon the accuracy of the history collected that was indicated by the system's use of the collected history in generating an accurate differential diagnostic list for each patient.

Comparison of Time to Complete and Number of Questions

Both collection modes had the possibility of asking 209 questions. The GQAP branching questionnaire entry system required the patient to view an initial 21 screens that contained a total of 89 questions at the noun level. Depending upon the answers to those questions the questionnaire branched to hierarchically lower questions. By the end of the questioning process, the questionnaire had stored answers to all possible questions. Answers to the questions were entered by the patient, or generated by no user screens within the questionnaire that entered "no" answers to questions hierarchically beneath questions to which the patient had previously entered a "no" answer.

In contrast, the DDA system required the patient to view a single initial screen containing six key questions. Based upon the answers to these questions the system, using a Bayesian scoring algorithm, evaluated the disease frames within the system in order to select a new set of questions, which were then submitted to the patient. This process continued until the patient had been queried for all relevant data and/or the total probability of the residual diagnoses had been reduced below the selected threshold. This process did not require an answer to be generated for all possible questions contained within the system.

The configuration of both history collection modes, as described above, was shown to greatly affect the amount of time and number of questions involved for each history collection mode to complete a history. Table 4 lists the variables that were used for comparison.

The patients who used the GQAP history collection mode took a mean time of 19.8 $\pm$ 6.6 (mean $\pm$ SD) minutes to complete their history collection session, whereas those using the DDA history program completed their session in 8.2 $\pm$ 5.8 minutes. The t-test calculated that a statistically significant ($p < .001$) difference existed between the amount of time each history collection mode took to complete a history collection session.

With the possibility of asking 209 total questions, GQAP asked (stored) a mean of 201.4 $\pm$ 7.1 total questions during a history collection session while DDA asked (stored) a mean of 63.4 $\pm$ 31.6 total questions. The category of total questions was subdivided into two additional categories: 1) the number of questions answered by the patient and 2) the number of questions that were answered by the program itself. While using the GQAP history collection mode, patients themselves answered a mean of 137.0 $\pm$ 19.1 questions with the program answering 64.4 $\pm$ 14.4 on its own. During DDA history collection sessions, patients answered a mean of 48.3 $\pm$ 29.0 questions while the program itself answered a mean of 15.1 $\pm$ 7.5 questions. Each of the above categories (total number of

Table 4. Variables from the QSTN and Query patient populations that were used as t-test variables for comparison of the two history collection methods.

Variable	QSTN (Mean ± SD)	Query (Mean ± SD)
Observations	24	52
Minutes*	19.8 ± 6.6	8.2 ± 5.8
Questions Answered Yes*	45.0 ± 17.5	22.1 ± 17.6
Questions Answered No*	92.0 ± 7.0	26.2 ± 13.8
Total Questions Asked*	201.4 ± 7.1	63.4 ± 31.6
Answered by Patient*	137.0 ± 19.1	48.3 ± 29.0
Answered by Program*	64.4 ± 14.4	15.1 ± 7.5

*Variables for which the t-test calculated the difference in means to be statistically significant
($p < .001$).

questions, questions answered by the patient, and questions answered by the program) for both history collection modes were compared using the T-Test. The T-Test calculated the difference in means in each category to be statistically significant ($p < .001$).

The category of questions answered by the patient was further subdivided into two categories; 1) questions answered "yes" and 2) questions answered "no". As patients completed their history using GQAP, they answered "yes" to a mean of 45.0 ± 17.5 questions and answered "no" to a mean 92.0 ± 7.0 questions. In contrast, those who had their histories collected by DDA answered "yes" to a mean of 22.1 ± 17.6 questions and "no" to a mean of 26.2 ± 13.8 questions. Use of the t-test again to compare these two categories showed the difference in means between the two history collection modes (GQAP and DDA) to be statistically significant ($p < .001$).

Comparison of Accuracy

To test the effectiveness of each history collection technique, the accuracy of the history collected was determined overall for each collection mode. The accuracy of the history itself was considered to be reflected in the accuracy of the five member differential diagnostic list, listed out for each patient after the respective system had gathered his/her history and then used that history as input to the knowledge base in order to generate diagnostic decisions.

The five member differential diagnostic list generated for each patient was compared to the admit and discharge diagnoses that were stored on the patient's file. The diagnostic list was examined to determine if it included the patient's admitting diagnosis, principal discharge diagnosis or any of the patient's pulmonary discharge diagnoses. Table 5 lists the admitting diagnosis, principal discharge diagnosis and any other pulmonary discharge diagnoses that were captured in the five member differential diagnostic lists generated by both the GQAP and DDA history collection modes.

In the group of patients who used the GQAP hierarchical questionnaire to enter their histories, 75% of the time (15 of 20) their admitting diagnosis was correctly identified in the differential diagnostic list. Those patients who used the DDA computer-directed history had their admitting diagnosis captured by the differential diagnosis list 86% (19 of 22) of the time. The chi-square test showed that there was a nonsignificant change in accuracy ($p < 0.59$).

For the patients whose history was collected using GQAP, the principal discharge diagnosis listed in the patient's file was captured 70% of the time (14 of 20) in the differential diagnostic list. The principal discharge diagnosis was identified 76% of the time (22 of 29) in the group of patients who used the DDA collection mode. The chi-

Table 5. Diagnoses captured by the QSTN and Query history collection modes used to compare the diagnostic accuracy between both history collection techniques.

Diagnosis Captured	QSTN	Query
Admitting Diagnosis	15/20 (75%)	19/22 (86%)
Principal Discharge Diagnosis	14/20 (70%)	22/29 (76%)
Any Pulmonary Discharge Diagnosis	29/44 (66%)	42/62 (68%)

Using the chi-square test to compare the diagnostic accuracy between QSTN and Query, in the three areas listed, showed no statistical significance in accuracy exists between these two methods of history collection.

square test calculated that no statistical significant difference in accuracy exists between the two methods for identifying the principal discharge diagnosis ($p < 0.915$).

Any pulmonary discharge diagnosis found in the patient's file was correctly identified in the differential diagnostic list 66% of the time (29 of 44) for those patients with GQAP collected histories. Patients who used DDA had any of their pulmonary discharge diagnoses captured in the differential list 68% of the time (42 of 62). The chi-square test showed that the change in accuracy between these two history collection methods for identifying any pulmonary discharge diagnosis was also nonsignificant ($p < 1.00$).

Results of Updating X-ray Statistics for Use in Sector Logic

This section contains the evaluation of the revised radiology statistics. Included in the evaluation is a comparison between the original and revised statistics (sensitivities and specificities), and a comparison of the accuracy of the differential lists produced using the original frames and those produced using the revised frames.

Comparison of Original Estimated Statistics to Revised Calculated Statistics

Table 6 contains the mean $\pm$ SD for the sensitivities and specificities in both the original and revised pulmonary disease supporting data frames. Only 10 of the 29 disease frames had six or more cases that could be used for evaluation in the training set in order to generate sensitivities. These 10 disease frames contained the 20 paired observations that were used to analyze changes in the sensitivities between the original and revised frames. The difference in means for the sensitivities, using the paired t-test, is considered statistically significant ($p< 0.05$).

New specificities were derived for all 29 disease frames using the training set. There were 78 paired observations used to analyze the change in specificities between the original and revised frames. The difference in means of the specificities in the two groups (original vs. revised), again using the paired t-test, is statistically significant ($p < 0.001$).

Comparison of Accuracy

To determine the effects of revising the frame statistics, both groups of frames (original and revised) were evaluated according to the effect they had upon the accuracy of he diagnostic system. The accuracy of each group of frames was based upon the ability of the system to include in the patients' differential diagnostic lists their respective discharge diagnoses (including No Pulmonary Disease), their discharge diagnoses for pulmonary diseases only, and their primary pulmonary discharge diagnosis.

Table 6. Original estimated statistics compared to revised calculated statistics for sensitivities and specificities in the pulmonary disease supporting data frames.

	Original (Mean ± SD)	Revised (Mean ± SD)
20 Paired Observations		
Sensitivity*	0.474 ± 0.264	0.367 ± 0.262
78 Paired Observations		
Specificity**	0.896 ± 0.129	0.941 ± 0.096

*$p < 0.05$ using paired t-test

**$p < 0.001$ using paired t-test

Both the training set and the test set were used in the evaluation. The training set, which had been utilized for the revision of the statistics, was used, even though it would show a biased change in accuracy because of its use in the revision, to determine exactly what effect the revised statistics had upon the accuracy of that group of patients' diagnostic lists. The test set was considered to be the gold standard for evaluating any changes made in the diagnostic accuracy associated with the use of the revised frames.

Table 7 lists the discharge diagnoses for all diseases, including No Pulmonary Disease, which were captured in the differential diagnostic lists of the patients in the test set. There was a possibility of capturing 110 total diseases in the test set. When only chest X-ray data were accessed by the system, the accuracy improved from 55% (original frames) to 87% (revised frames) for the discharge diagnosis to be included in the five member list, and from 41% to 56% for the discharge diagnosis to be ranked number one on the list. The change in accuracy for both these classifications was determined to be statistically significant using McNemar's Test.

When both history and chest X-ray data were accessed by the system, the accuracy for including the discharge diagnosis within the five member list went from 87% to 90%. The system's ability to rank the discharge diagnosis as number one fell from 66% to 64%. Both these changes in accuracy between the original and revised frames were considered to be nonsignificant (McNemar's Test).

Table 8 lists the discharge diagnoses for all diseases, including No Pulmonary Disease, which were captured in the differential diagnostic lists of the patients within the training set. There was a total of 548 discharge diagnoses within the training set. The lists created using chest X-ray data alone improved their accuracy from 64% (original) to 89% (revised) for including the discharge diagnosis within the five member list. The system's ability to rank the discharge diagnosis number one on the list improved from 56% to 68%. The change inaccuracy for both groups between the original and revised frames was considered to be statistically significant (McNemar's Test).

The lists created by accessing both history and chest X-ray data improved their accuracy from 90% to 92% when including the discharge diagnosis in the five member list ($0.05 \leq p < 0.1$, which indicates a trend toward statistical significance). The accuracy stayed the same between the original and revised frames for ranking the discharge diagnosis number one (70%).

Table 9 lists the discharge diagnoses for pulmonary diseases only, which were captured in the differential diagnostic lists in the test set. There was a possibility of capturing 58 pulmonary diseases in the test set. When the system accessed chest X-ray findings only in order to create the diagnostic lists, the accuracy of the system improved

Table 7. Discharge diagnoses for all diseases, including No Pulmonary Disease, captured in the differential diagnosis lists (test set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	61/110 (55%)	96/110 (87%)*
Ranked Number 1	45/110 (41%)	62/110 (56%)**
History + X-ray Results		
Included in 5 Member List	96/110 (87%)	99/110 (90%)
Ranked Number 1	73/110 (66%)	70/110 (64%)

*$p < 0.001$ using McNemar's Test

**$p < 0.005$ using McNemar's Test

Table 8. Discharge diagnoses for all diseases, including No Pulmonary Disease, captured in the differential diagnosis lists (training set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	353/548 (64%)	485/548 (89%)*
Ranked Number 1	305/548 (56%)	371/548 (68%)*
History + X-ray Results		
Included in 5 Member List	494/548 (90%)	504/548 (92%)**
Ranked Number 1	383/548 (70%)	384/548 (70%)

*$p < 0.001$ using McNemar's Test

** $0.05 \leq p < 0.1$ indicating a trend toward statistical significance

Table 9. Discharge diagnoses for all pulmonary diseases captured in the differential diagnosis lists (test set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	10/58 (17%)	44/58 (76%)*
Ranked Number 1	3/58 (5%)	23/58 (40%)*
History + X-ray Results		
Included in 5 Member List	46/58 (79%)	49/58 (84%)
Ranked Number 1	28/58 (48%)	29/58 (50%)

*$p < 0.001$ using McNemar's Test

from 17% (original) to 76% (revised) for the inclusion of the pulmonary disease within the five member list. The accuracy went from 5% to 40% for ranking the pulmonary discharge diagnosis as number one in the list. Both of the above differences in accuracy between the original and the revised frames were significant using McNemar's Test.

When both history and chest X-ray data were used by the system to create the diagnostic lists, the system improved its accuracy to include the pulmonary discharge diagnosis within the five member list from 79% to 84%. The ability of the system to rank the pulmonary discharge diagnosis as number one changed from 48% to 50%. These changes in accuracy were not considered statistically significant (McNemar's Test).

Table 10 is a list of the discharge diagnoses for all pulmonary diseases that were captured in the differential diagnostic lists within the training set. There was a total of 220 possible pulmonary discharge diseases contained in the training set. For those lists generated using chest X-ray data alone, the system was able to improve its accuracy of including the pulmonary discharge diagnosis within the five member list from 13% (original) to 71% (revised). It was able to rank the pulmonary disease as number one with an improvement in accuracy from 2% to 33%. Both of these changes in accuracy between the original and revised frames were significant (McNemar's Test).

As the system looked at both history and chest X-ray data to produce the diagnostic lists, the accuracy went from 78% to 83% for the inclusion of the pulmonary disease within the five member list. The system was able to rank the pulmonary discharge diagnosis as number one with an improvement from 37% to 42%. The difference between the original and revised frames for both these groups was also considered significant using McNemar's Test.

The final area in which the accuracy was compared between the original and revised logic frames was in the system's ability to capture, for inclusion in and to rank as number one in the diagnostic list, the patient's primary pulmonary discharge diagnosis. Table 11 contains the primary pulmonary discharge diagnoses that were captured in the lists in the test set. There was a total of 20 primary pulmonary discharge diagnoses in the test set. The system, using only access to chest X-ray finding, improved in accuracy from 30% (original) to 80% (revised) for including the primary pulmonary discharge diagnosis within the five member list. Its ability to rank the primary pulmonary discharge diagnosis as number one improved from 5% to 40%. Both these changes in accuracy were determined to be statistically significant (McNemar's).

When the system accessed both history and chest X-ray findings, the accuracy with which it included the primary pulmonary discharge diagnosis in the five member list did not change (95%). It improved its ability to rank the primary pulmonary discharge diagnosis

Table 10. Discharge diagnoses for all pulmonary diseases captured in the differential diagnosis lists (training set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	28/220 (13%)	157/220 (71%)*
Ranked Number 1	5/220 (2%)	72/220 (33%)*
History + X-ray Results		
Included in 5 Member List	172/220 (78%)	183/220 (83%)**
Ranked Number 1	81/220 (37%)	93/220 (42%)***

*$p < 0.001$ using McNemar's Test

**$p < 0.05$ using McNemar's Test

***$p < 0.025$ using McNemar's Test

Table 11. Primary pulmonary discharge diagnoses captured in the differential diagnosis lists (test set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	6/20 (30%)	16/20 (80%)*
Ranked Number 1	1/20 (5%)	8/20 (40%)*
History + X-ray Results		
Included in 5 Member List	19/20 (95%)	19/20 (95%)
Ranked Number 1	12/20 (60%)	14/20 (70%)

*$p < 0.05$ using McNemar's

as number one in the list from 60% to 70%. Neither of the above changes in accuracy was considered significant.

Table 12 lists the primary pulmonary discharge diagnoses that were captured in the differential diagnostic lists for the patients in the training set. The training set contained a total of 93 primary pulmonary discharge diagnoses. The lists that were generated while accessing only chest X-ray findings improved in accuracy from 13% to 71% for the inclusion of the diagnosis in the five member list. The system ranked the diagnosis as number one with an improved accuracy from 2% to 41%. Both these differences in accuracy between the original and revised frames are statistically significant using McNemar's Test.

When the system was able to use both history and X-ray data, the accuracy of the system to include the diagnosis in the five member list went from 83% to 91%. Its ability to rank the primary pulmonary discharge diagnosis as number one increased from 41% to 52%. The difference between the original and revised systems for both these groups was also considered significant (McNemar's).

Table 12. Primary pulmonary discharge diagnoses captured in the differential diagnosis lists (training set).

	Original Frames	Revised Frames
X-ray Results Only		
Included in 5 Member List	12/93 (13%)	66/93 (71%)*
Ranked Number 1	2/93 (2%)	38/93 (41%)*
History + X-ray Results		
Included in 5 Member List	77/93 (83%)	85/93 (91%)**
Ranked Number 1	38/93 (41%)	48/93 (52%)***

*$p < 0.001$ using McNemar's Test

**$p < 0.05$ using McNemar's Test

***$p < 0.005$ using McNemar's Test

DISCUSSION AND CONCLUSION

GQAP Verses DDA

Discussion

Patient history is an important source of information, which is time consuming and often difficult to capture. Because of difficulties in routinely capturing patient histories, these data are often lacking from most hospital information systems. Patient history provides a wealth of information that is necessary in establishing a differential diagnostic list for the patient, in most clinical settings, which will direct the work-up and therapy for that particular patient. This project focused upon the collection of patient histories with patients entering their history data into the HIS via direct interaction with a terminal at the patient's bedside.

In this study two different computerized data gathering techniques were compared, GQAP, a hierarchical questionnaire that uses a conditional branching set of fixed frames, and DDA, a decision-driven data acquisition system that revolves around the concept of hypothesis directed questioning. The goal was to determine if there existed a measurable difference in 1) time to complete, 2) number of questions asked or 3) accuracy of decision made from the entered data.

In comparing the two modes in the areas of time to complete and number of questions asked, a statistically significant difference was found to exist between them. GQAP took a mean time of 19.8 ± 6.6 minutes to complete and asked a mean of 201.4 ± 7.1 total questions. Of these 201.4 ± 7.1 total questions asked, a mean of 64.4 ± 14.4 was inferred as "no" answers by GQAP. A mean of 137.0 ± 19.1 questions was answered by the patient. Of these questions answered by the patient, a mean of 45.0 ± 17.5 was answered "yes" with a mean of 92.0 ± 7.0 being answered "no." DDA took a mean time of 8.2 ± 5.8 minutes to complete a history collection session and asked a mean of 63.4 ± 31.6 total questions. A mean of 15.1 ± 7.5 was inferred as "no" answers by DDA. A mean of 48.3 ± 29.0 questions was answered by the patient. Of these, a mean of 22.1 ± 17.6 was answered "yes" with a mean of 26.2 ± 13.8 being answered "no."

DDA was able to markedly reduce the number of questions asked in all the categories that were examined. This in turn reduced the time it took to complete a history using the DDA history collection mode. The reduction in the number of questions asked to those patients using the DDA history collection mode is due in a large part to the smaller amount of "no" answers collected by the program, both those that were inferred by the DDA program itself and those that were answered "no" by the patient. The process of decision-driven data acquisition is responsible for the bulk of reduction in the questioning process. DDA is designed to request only that data directly needed for the exploration of specific hypotheses about a given patient's condition. Therefore, data collection is limited by the decisions under consideration. Theoretically this technique has the advantage of gathering data adequate to determine accurate diagnoses without exhaustive approaches to questioning. This not only reduces the amount of time a patient needs to spend to complete a history collection session, but also helps in eliminating questions that are inapplicable to the patient's current health problem(s).

Although DDA asked fewer questions of each patient than did GQAP, when the two data collection modes were compared based upon accuracy, it was shown that there existed no statistical difference between them. The differential diagnostic lists that were created using history data stored by both collection modes were compared to the admit and discharge diagnoses stored on the patient's file. QGAP captured 75% of the possible admitting diagnoses, 70% of the principal discharge diagnoses, and 66% of any pulmonary discharge diagnoses that were stored on the files of patients using the GQAP history collection program.The patients who used the DDA history collection program had their admitting diagnoses captured 86% of the time, their principal discharge diagnoses were captured 76% of the time, and any pulmonary discharge diagnoses they had stored on their file were captured 68% of the time.

During the implementation and evaluation of this study several problems became apparent not only with the history collection process but also with the data base as it affected the accuracy of the diagnostic lists generated. During the history collection phase of the study, one of the daily problems dealt with was the difficulty of collecting all possible histories of patients that were considered to be part of defined patient population. Since the history collection was not part of the routine hospital stay, it was hard to catch all possible patients after they had been admitted before the 48 hour time limit was up. It was sometimes difficult to get the patients to participate in the study, as they felt it was just another bothersome unnecessary thing happening during their hospital stay. The condition of the patients also affected their willingness and ability to participated. Some patients had to have the questions read to them, and others had to have help entering their histories into

the computer. Some had to have help with both. Questions were reworded before they were used in the history collection programs, but even after coming up with what was thought to be the best formats for the questions, some patients still had trouble determining what information was really being sought by the question. They also had difficulty in being able to answer some question with a "yes" or "no" answer. Some felt that having the option of answering "unknown" would have been useful to them. Many patients felt frustrated and discouraged, especially when using the GQAP history collection program, at the length of the questioning process. Sometime it seemed to them that the questioning process would never be over. As their frustration increased it was felt that their attention to the questions being asked them decreased and the accuracy of their history was possibly decreased as a result.

During the evaluation of this project several factors were recognized that worked to reduce the accuracy of the diagnostic lists that were generated using history data as input. The first is the fact that the knowledge base did not always have a sector for the disease the patient was presenting. Even though the patients were screened to ensure they belonged to the group of cardio-pulmonary physicians who agreed to let us use their patients in the study, it was found that occasionally in that narrow medical speciality a sector had not yet been created for a particular, usually less common, cardio-pulmonary disease. This would in return reflect upon the accuracy of the history collected, in that even though the history collected had been adequate to accurately diagnose the disease present the differential diagnostic list generated for that patient would not have included the disease and the list would have been considered inaccurate because of the miss only because the disease did not exist in the knowledge base. This type of inaccuracy will slowly improve as new logic frames are created for diseases that are not already included in the knowledge base.

The second factor that contributed to inaccuracies in the system is the fact that the discharge diagnoses, that were used as the standard for comparison, are entered into the patients' files by the medical records department after the patients are discharged. The history collection programs were designed to identify diseases which the patients actively had on their admission to the hospital, while sometimes the discharge diagnoses do not always reflect the patients' primary diseases and may reflect complications occurring from procedures performed in the hospital after admission or hospital acquired diseases. This results in the differential diagnostic lists created from the collected history being compared to discharge diagnoses that may or may not reflect those diseases present at the time of admission. If the discharge diagnoses do not accurately reflect diseases present on admission this comparison could lead to apparent misses as the diagnostic system

produced what appears as an erroneous diagnosis. This in turn would falsely reflect that the history collected was inaccurate.

The third factor that possibly reduced the accuracy of the generated diagnostic lists, is the accuracy of the sensitivities and specificities incorporated in the Bayesian model used in the disease logic frames along with the logic itself. As improvements are made in the logic of the disease frames and statistics are updated, errors in accuracy due to these will decrease as the system itself becomes more accurate in identifying each disease. As was shown in an earlier study, when the Bayesian statistics are derived from an actual clinical data base, instead of relying upon estimates, the system's performance is greatly increased, but is still not up to 100%. These newly revised statistics for the history frames were used in this study. As revision using clinical data bases continues the system will only continue to improve upon its ability to identify diseases.

Conclusion

This study has shown that the DDA history collection mode appears to be a better approach in the collection of patient histories than the GQAP history collection mode. The decision-driven data acquisition tools greatly reduced the number of questions each patient was required to answer and thus reduced the amount of time each patient had to spend to complete a history collection session. The reduction in questions asked of the patients did not affect the diagnostic accuracy (in some cases it appeared as an improvement in accuracy had occurred).

It was shown that using a hierarchical branching questionnaire (GQAP) is not an efficient way to enter data (patient histories) into the HIS. Branching questionnaires appear to work if they are limited to gathering data in a narrow area of medicine, but as the group of diagnostic modules grow, thereby increasing to possible numbers of questions presentable to each patient, the importance of a collection programs such as DDA becomes apparent in its ability to "intelligently" question patients. DDA's use of a set of diagnostic hypotheses to condition the questioning process appears to be effective in capturing a relevant history and in reducing the burden to the patient.

History collection can now be viewed as more easily obtainable and once captured can be made available for use throughout the hospital information system. As computer terminals become more abundant throughout the hospital, being installed in patient rooms, it would be exciting to see patient history collection by the system becoming a routine admit procedure. These data could be an important diagnostic tool for physicians. It would also enlarge the clinical data base and make headway for new research and development on the system.

As the technique for history collection becomes refined, it may become useful for collection of other clinical data. One of the next areas of development on the system will involve the physical exam being entered by the physician at the bedside. DDA would be used to reduce the respondent's input chores while gathering that information necessary for discrimination among likely diagnoses. As the statistics for those logic frames have been revised, as stated elsewhere in this thesis, the next step is to begin collection of the physical exam at the patient bedside. This will further enhance the system and its medical decision making abilities.

Data collected using this technique, both the history and physical exam, could then be used to develop new avenues in the medical decision making area. For example, the data collected could be used as input into a system whose goal is defining the treatment and medical therapy necessary for individual patients. The knowledge base, as it grows and becomes more extensive, along with the data collected could be the source for an excellent educational tool. The patients, once finishing a history collection session, could then learn from the system what their medical problems are and what they could do in order to help with their recovery. The data and data base could also be used for health care professional personnel education. There seems to many areas where data collection and decision making could be of great use in the medical field. They will just have to be explored and developed as the data collection techniques become more and more refined.

Revision of X-ray Statistics

Discussion

In order for the next phase of data collection to begin, entering a physical exam at the patient's bedside, it was felt that the statistics used in the Bayesian system for generation of a differential diagnostic list needed to be revised. The original Bayesian statistics, sensitivities and specificities for chest X-ray findings in each disease were initially estimated by a group of medical experts. These statistics were revised using a clinical data base and analyzed to determine what effect this revision would have on the functionality of the diagnostics system. Were they were an improvement over the originally estimated ones?

In comparing the estimated statistics to the derived statistics, it was seen that both over- and underestimations had been made by the medical experts. On the whole, the sensitivities had been overestimated by the medical experts, the mean of the estimated sensitivities being 0.474 ± 0.264 while the mean of the revised sensitivities was 0.367 ± 0.262. Overall the specificities had been underestimated by the medical experts. The mean

of the estimated specificities was 0.896 ± 0.129 with the mean of the revised specificities being 0.941 ± 0.096.

To determine the significance of these estimation errors, their effect upon the accuracy of the diagnostic system was evaluated. Following the revision of the diagnostic logic, the diagnostic results of the original frames were compared to those of the revised frames for both the training and test set. In both the training and test set a statistically significant improvement in accuracy occurred in the generation of a differential diagnostic list when statistics derived from the data base were substituted for the estimated values.

The differential diagnostic lists were evaluated in both the training set and test set for the following features: 1) the rate for which the correct "true" diagnosis was included in the five member list, 2) the rate for which the correct "true" diagnosis was ranked number one in the list, 3) the rate for which the primary discharge diagnosis was included in the list, 4) the rate for which the primary discharge diagnosis was ranked number one, 5) the rate at which the total pulmonary diseases were captured in the list, and 6) the rate for which the existing pulmonary disease(s) ranked number one. It was shown that in every area there was a statistically significant improvement in accuracy of the diagnostic system when the revised statistics were used and the diagnostic frames were limited to access only chest X-ray data.

When the revised statistics were used and the frames were allowed to access both history and chest X-ray data there were: nine cases in which an improvement in accuracy was made, with five of those cases being statistically significant; one case in which the accuracy dropped 2%; and two cases in which the accuracy stayed the same. It is felt that since the history frames were already using revised statistics and were accessed by the system before the physical exam frames, which contained the revised statistics for chest X-ray findings, that the history frame diagnostic results worked to diminish the effect of the revised chest X-ray statistics. This is mainly because the a posteriori probability created by the history frame for each disease was used as the a priori probability for the physical exam logic frame for that disease.

In retrospect, one problem with this study was the limited size of the data base that was used. The training set that was used for the revision of the statistics consisted of only 535 patients. Of the 29 disease frames needing revision of statistics, only 10 disease frames were represented by sufficient patients (six or more) so that both revised sensitivities and specificities could be generated. This problem will slowly improve as the clinical data base is added to with new patients manifesting more of the represented diseases. In order to keep the system improving, the data base needs to be added to and

used to created new revised statistics for the use of upgrading the diagnostic frames on a regular basis.

Conclusion

We have shown in this analysis that the diagnostic system using the statistical values derived from a clinical data base performs significantly better than the original system that utilized estimated Bayesian statistics. The use of a clinical data base -- representative of the types of patients for whom the diagnostic system is designed -- to derive the statistics used in a Bayesian approach to diagnostic decision making, is very important in influencing the accuracy of the diagnostic system.

As statistics derived from real clinical data bases have been shown to perform more accurately, the next area of development would be to create an "intelligent" self-learning system. This system would be able to review the clinical data base in order to update and revise the diagnostic logic on a regular basis as the clinical data base changes with growth.

REFERENCES

1. Lilford RJ, Glyn-Evans D, Chard T (1983): The use of patient-interactive microcomputer system to obtain histories in an infertility and gynecologic endocrinology clinic. Am J Obstet Gynecol 146(4):374-379.

2. Weed LL (1968): Medical records that guide and teach. New Engl J Med 278:593-600.

3. Coombs GJ, Murray WR, Krahn DW (1970): Automated medical histories: factors determining patient performance. Comp Biomed Res 3 178-181.

4. Slack WV, Hicks GP, Reed CE, Van Cura LJ (1966): A computer-based medical-history system. New Engl J Med 274(4):194-198.

5. Houziaux M-O, Lefebvre PJ (1986): Historical and methodological aspects of computer-assisted medical history-taking. Med Inform 11(2):129-143.

6. Teach RL, Shortliffe EH (1981): An analysis of physician attitudes regarding computer-based clinical consultation systems. Comp Biomed Res 14 542-558.

7. Fox J, Alvey P (1983): Computer assisted medical decision making. British Med J 287:742-746.

8. McDonald CJ (1976): Protocol-based computer reminders, the quality of care and the non-perfectability of man. New Engl J Med 295(24):1351-1355.

9. McDonald CJ, Hui SL, Smith DM, Tierney WM, Cohen SJ, Weinberger M, McCabe P (1984): Reminders to physicians from an introspective computer medical record. Annals Intern Med 100:130-138.

10. Dombal FT, Leaper DJ, Horrocks JC, Standiland JR, McCann AP (1974): Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. British Med J 1:376-380.

11. Kiely JM, Juergens JL, Hisey BI, Williams PE (1963): A computer-based medical record. JAMA 205(8):571-576.

12. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB (1976): Towards the simulation of clinical cognition: taking a present illness by computer. Am J Med 60:981-996.

13. Lilford RJ, Bingham P, Fawdry R, Setchell M, Chard T (1983): The development of on-line history-taking systems in antenatal care. Meth Inform Med 22:189-197.

14. Croft DJ (1972): Is computerized diagnosis possible? Comp Biomed Res 5:351-367.

15. Gorry GA, Barnett GO (1968): Experience with a model of sequential diagnosis. Comp Biomed Res 1:490-507.

16. Shortliffe EH., Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN (1975): Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comp Biomed Res 8:303-320.

17. Engle RL., Flehinger BJ, Allen S, Friedman R, Lipkin M, Davis BJ, Leveridge LL (1976): HEME: a computer aid to diagnosis of hematologic disease. Bull N Y Acad Med 52(5):584-599.

18. Fieschi M, Joubert M, Fieschi D, Roux M (1982): SPHINX - a system for computer-aided diagnosis. Meth Inform Med 21:143-148.

19. Miller RA, Pople HE, Myers JD (1982): INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. New Engl J Med 307(8):468-476.

20. Warner HR (1979): Computer-assisted medical decision-making. Academic Press New York.

21. Warner HR, Haug PJ, Pryor TA (Submitted March 1987): Report on the grant "Can an 'expert system' facilitate medical data entry?" NIH Grant HS 03810.

22. Pryor TA, Gardner RM, Warner HR (1983): The HELP system. J Med Systems 7(2):87-102.

23. Gerard MJ (Dec 1984): Computerized automated pulmonary disease diagnosis. Thesis, Department of Medical Biophysics and Computing. University of Utah.

24. Pryor TA, Warner HR, Gardner RM, et al. (1988): The HELP system development tools. In: Orthner H and Blum B (eds.), Methods for Developing Clinical Information Systems. New York: Springer Verlag Inc.

25. Warner HR, Rutherford BD, Houtchens B (1972): A sequential bayesean approach to history taking and diagnosis. Comp Biomed Res 5:256-262.

26. Haug PJ, Warner HR, CLayton PD, Schmidt CD, Pearl JE, Farney RF, Crapo RO, Tocino I, Morrison WJ, Fredrick PR (1987): A decision-driven system to collect the patient history. Comp Biomed Res 20:193-207.

27. Haug PJ, Warner HR, Clayton PD, Schmidt CD, Pearl JE, Farney RJ (1986): A computer-directed patient history: functional overview and initial experience. MEDINFO-86:849-852.

28. Haug PJ, Warner HR (1984): Decision-driven acquisition of qualitative data. Proceedings of the Eighth Annual Symposium on Computer Applications in Medical Care, Washington, DC 189-192.

29. Warner HR ,Haug PJ (1983): Medical data acquisition using an intelligent machine. MEDINFO-83:582-584.

30. Shelton P: Analysis of the information content of medical data using a frame-based medical diagnostic system. Unpublished Manuscript, Department of Medical Informatics. University of Utah.

31. Haug PJ, Clayton PD, Shelton P, Rich T, Tocino I, Frederick PR, Crapo RO, Morrision J (1987): Revision of diagnostic logic using a clinical data base. Proceedings AAMSI Conference, San Francisco 1987.